

# Ein Simpson-Paradoxon bei Covid-19-Todesfallraten

NORBERT HENZE, KARLSRUHE

**Zusammenfassung:** Ein aktueller Datensatz über den Anteil der Todesfälle unter allen bestätigten Covid-19-Infektionen zeigt, dass die Überlebenschancen, bezogen auf die Gesamtbevölkerung, in China besser sind als in Italien. Schaut man jedoch genauer hin und vergleicht die Todesfallraten innerhalb jeder der Altersgruppen 0-9 Jahre, 10-19 Jahre usw., so ergibt sich in allen Fällen ein umgekehrtes Bild. Dieses auf den ersten Blick überraschende Phänomen ist zum Teil den unterschiedlichen Alterspyramiden in beiden Ländern geschuldet. Hier offenbart sich ein weiteres, unterrichtsrelevantes Beispiel für das Auftreten des Simpson-Paradoxons.

## 1 Der Datensatz

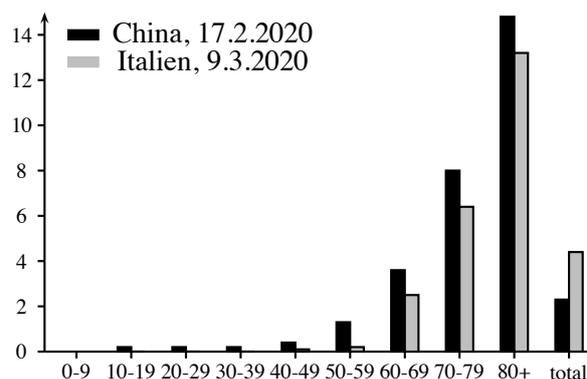
Die aktuelle Coronavirus-Pandemie geht auf das SARS-CoV-2-Virus zurück. Dieses Virus verursacht die ansteckende Atemwegserkrankung Covid-19. Ein wichtiger Indikator ist hierbei die im Folgenden mit TFR abgekürzte Todesfallrate (engl.: case fatality rate, CFR). Die TFR ist der Prozentsatz tödlich endender Fälle unter allen bestätigten Covid-19-Infektionen. Da die TFR stark vom Alter abhängt, wobei Patienten mit höherem Alter tendenziell einer höheren TFR ausgesetzt sind, werden Todesfallraten häufig auch getrennt nach Altersgruppen veröffentlicht. Die von Kügelgen et al. (2020) entnommenen Daten von Tab. 1 zeigen nach Altersgruppen klassifizierte Todesfallraten aus China (Stand: 17.2.2020) und Italien (Stand: 9.3.2020).

Alter	0-9	10-19	20-29	30-39	40-49
Italien	<b>0%</b> <small>(0/43)</small>	<b>0%</b> <small>(0/85)</small>	<b>0%</b> <small>(0/296)</small>	<b>0%</b> <small>(0/470)</small>	<b>0.1%</b> <small>(1/891)</small>
China	<b>0%</b> <small>(0/0)</small>	<b>0.2%</b> <small>(1/549)</small>	<b>0.2%</b> <small>(7/3.619)</small>	<b>0.2%</b> <small>(18/7.600)</small>	<b>0.4%</b> <small>(38/8.571)</small>
Alter	50-59	60-69	70-79	≥ 80	total
Italien	<b>0.2%</b> <small>(3/1.453)</small>	<b>2.5%</b> <small>(37/1.471)</small>	<b>6.4%</b> <small>(114/1.785)</small>	<b>13.2%</b> <small>(202/1.532)</small>	<b>4.4%</b> <small>(357/8.026)</small>
China	<b>1.3%</b> <small>(130/10.008)</small>	<b>3.6%</b> <small>(309/8.583)</small>	<b>8%</b> <small>(312/3.918)</small>	<b>14.8%</b> <small>(208/1.408)</small>	<b>2.3%</b> <small>(1.023/44.672)</small>

**Tab. 1:** Todesfallraten nach Altersgruppen aufgeschlüsselt sowie aggregiert

Dabei sind die im Ländervergleich jeweils geringeren Raten durch Fettdruck hervorgehoben. Ein in Kleinschrift gesetzter Eintrag ( $x/y$ ) bedeutet  $x$  Todesfälle von insgesamt  $y$  erfassten Corona-Infektionen. Die Tabelle zeigt unter „total“ auch die

über alle Altersgruppen aggregierten Werte. Ins Auge springt, dass über alle erfassten Fälle gesehen die TFR in Italien höher ist als in China. Innerhalb jeder Altersgruppe stellt sich die Situation jedoch genau andersherum dar. Abbildung 1 veranschaulicht dieses Phänomen in Form einer Grafik.



**Abb. 1:** Todesfallraten (in Prozent) nach Altersgruppen klassifiziert sowie aggregiert

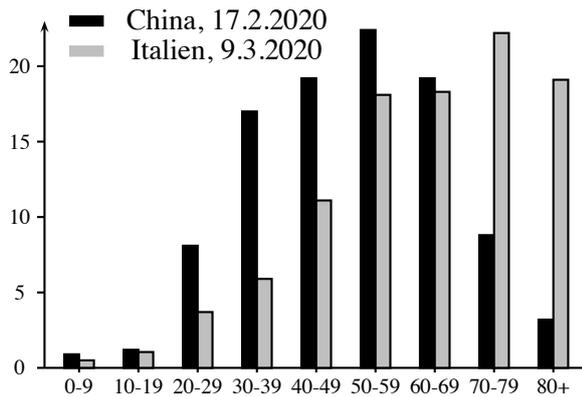
Dieses auf den ersten Blick überraschende Phänomen, dass ein Merkmal (hier: die TFR) beim Vergleich zweier Populationen 1 und 2 in Population 1 prozentual öfter auftritt als in 2, obwohl es bei Schichtung der Population (hier: in Altersklassen) innerhalb jeder Schicht prozentual seltener vorkommt, ist aufgrund der Arbeit Simpson (1951) nach dem britischen Statistiker Edward Hugh Simpson (1924–2019) als *Simpson-Paradoxon* bekannt. Zu diesem Paradoxon gibt es eine umfangreiche Literatur, siehe z.B. Blyth (1972), Engel und Sedlmeier (2010), Knapp (1985), Kocik (2001), Künzel (1991), Lesser und Pearl (2020), Meyer (1995), Motzer (2009), Newson (1991), Parliades and Perlman (2009), Reintjes et al. (2000), Samuels (1993), Wagner (1982) und Wardrop (1995).

Alter	0-9	10-19	20-29	30-39	40-49
Italien	0.5%	1.1%	3.7%	5.9%	11.1%
China	<b>0.9%</b>	<b>1.2%</b>	<b>8.1%</b>	<b>17.0%</b>	<b>19.2%</b>
Alter	50-59	60-69	70-79	≥ 80	
Italien	18.1%	18.3%	<b>22.2%</b>	<b>19.1%</b>	
China	<b>22.4%</b>	<b>19.2%</b>	8.8%	3.2%	

**Tab. 2:** Anteile der einzelnen Altersgruppen an allen bestätigten Covid-19-Fällen

Den Schlüssel zum Verständnis des kontraintuitiven Verhaltens der globalen Prozentzahlen im Vergleich

zu den Werten innerhalb der einzelnen Altersgruppen liefert ein Blick auf Tab. 2. Die dort angegebenen Werte sind die in Prozent angegebenen Anteile der einzelnen Altersgruppen an der Gesamtzahl aller bestätigten Covid-19-Infektionen. So gehörten etwa von den 44.672 insgesamt in China bestätigten Fällen 8.583 der Altersgruppe 60-69 an (siehe Tab. 1), was zu dem in der letzten Zeile von Tab. 2 aufgeführten Prozentsatz 19.2% führt. Abbildung 2 stellt die Daten von Tab. 2 grafisch dar.

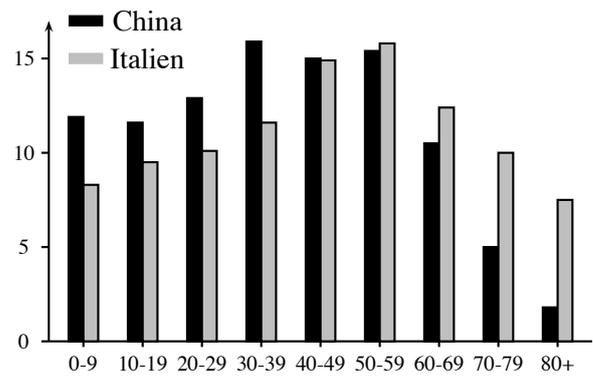


**Abb. 2:** Prozentanteile der Altersgruppen an allen bestätigter Infektionen

Offensichtlich differiert die Verteilung aller bestätigten Fälle auf die jeweiligen Altersgruppen erheblich zwischen beiden Ländern. So wurde in Italien im Vergleich zu China ein weit höherer Anteil von Fällen bei älteren Patienten festgestellt. Obwohl die Fälle sowohl in Italien als auch in China lokal sehr ungleich verteilt waren, zeigt ein Blick auf Tab. 3 und Abb. 3 (Daten aus von Kügelgen et al. (2020)), dass sich die Altersverteilungen in China und Italien deutlich unterscheiden. Insofern ist diese ungleiche Altersverteilung mit einer im Vergleich zu China im Durchschnitt älteren Bevölkerung in Italien hauptverantwortlich für das festgestellte Phänomen.

Alter	0-9	10-19	20-29	30-39	40-49
Italien	8.3%	9.5%	10.1%	11.6%	14.9%
China	<b>11.9%</b>	<b>11.6%</b>	<b>12.9%</b>	<b>15.9%</b>	<b>15%</b>
Alter	50-59	60-69	70-79	≥ 80	
Italien	<b>15.8%</b>	<b>12.4%</b>	<b>10%</b>	<b>7.5%</b>	
China	15.4%	10.5%	5%	1.8%	

**Tab. 3:** Altersverteilung der Bevölkerungen in Italien und China



**Abb. 3:** Altersverteilungen in China und Italien

## 2 Theoretischer Hintergrund

Stellen wir uns vor, alle 53.243 (= 44.672 + 8.571) insgesamt in China und Italien erfassten Covid-19-Fälle seien gedanklich in irgendeiner Weise durchnummeriert und als Menge  $\Omega$  modelliert. Wenn wir rein zufällig einen dieser Fälle herausgreifen, legen wir (ebenfalls gedanklich) die Gleichverteilung  $\mathbb{P}$  auf  $\Omega$  zugrunde, arbeiten also mit einem Laplace-Modell. Bezeichnen  $C$  und  $I$  die Teilmengen von  $\Omega$ , die sich auf die Fälle in Italien bzw. in China beziehen, so sind diese Ereignisse komplementär zueinander, d.h., es gelten  $I = \bar{C}$  und  $C = \bar{I}$ . Wir bilden jetzt eine Zerlegung der Menge  $\Omega$  in neun paarweise disjunkte Teilmengen  $K_1, \dots, K_9$ , nämlich (in dieser Reihenfolge) in die Altersklassen 0-9, 10-19, ..., 70-79 und 80+. Bezeichnet die Teilmenge  $S$  von  $\Omega$  alle tragischen Fälle, die die Covid-19-Infektion nicht überlebt haben, und schreiben wir allgemein

$$\mathbb{P}_C(B) := \mathbb{P}(B|C), \quad \mathbb{P}_I(B) := \mathbb{P}(B|I), \quad B \subseteq \Omega,$$

für die bedingten Wahrscheinlichkeiten eines Ereignisses  $B$  unter der Bedingung  $C$  bzw.  $I$ , so gelten

$$\mathbb{P}_C(S) = \frac{1.023}{44.672} \approx 0.023, \quad \mathbb{P}_I(S) = \frac{357}{8.026} \approx 0.044$$

und folglich  $\mathbb{P}_C(S) < \mathbb{P}_I(S)$ . Andererseits gelten (bis auf die Altersgruppe  $K_1$ , in der keine tödlich verlaufenden Fälle auftreten) die „umgekehrten Ungleichungen“

$$\mathbb{P}_C(S|K_j) > \mathbb{P}_I(S|K_j) \text{ für jedes } j = 2, \dots, 9.$$

So liest man etwa aus Tabelle 1

$$\begin{aligned} \mathbb{P}_C(S|K_7) &= \frac{309}{8.583} \approx 0.086, \\ \mathbb{P}_I(S|K_7) &= \frac{37}{1.471} \approx 0.025 \end{aligned}$$

ab. Wenn man sich vor Augen führt, dass sich die Wahrscheinlichkeiten  $\mathbb{P}_C(S)$  und  $\mathbb{P}_I(S)$  nach der Formel von der totalen Wahrscheinlichkeit (siehe z.B. Henze (2018), S.104) in der Form

$$\mathbb{P}_C(S) = \sum_{j=1}^9 \mathbb{P}_C(K_j) \mathbb{P}_C(S|K_j),$$

$$\mathbb{P}_I(S) = \sum_{j=1}^9 \mathbb{P}_I(K_j) \mathbb{P}_I(S|K_j)$$

als gewichtete Summen der bedingten Wahrscheinlichkeiten  $\mathbb{P}_C(S|K_j)$  bzw.  $\mathbb{P}_I(S|K_j)$  ergeben, wird klar, warum in diesem Fall das Simpson-Paradoxon auftritt. So ist etwa

$$\mathbb{P}_C(K_8) = \frac{3.918}{44.672} \approx 0.088, \quad \mathbb{P}_I(K_8) = \frac{1.785}{8.026} \approx 0.22$$

(vgl. die Prozentzahlen in Tab. 2 sowie Abb. 2). In diesem Fall tragen insbesondere die beiden im Vergleich zu China hohen Gewichte  $\mathbb{P}_I(K_8)$  und  $\mathbb{P}_I(K_9)$  der beiden obersten Altersgruppen maßgeblich zum Simpson-Paradoxon bei.

Gerade die Tatsache, dass Schülerinnen und Schüler massiv von der Corona-Pandemie betroffen sind, kann ein auf den ersten Blick überraschendes Phänomen, das sich auf diese Pandemie bezieht, zu wichtigen Diskussionen führen. Auf jeden Fall trägt es zur Datenkompetenz bei.

**Danksagung:** Ich danke dem Schulgutachter für wertvolle Hinweise.

## Literatur

- Blyth, C.R. (1972). Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Ass.* 67(338), 364–366.
- Engel, J., und Sedlmeier, P. (2010). Regression und Korrelation: Alles klar, oder voller Tücken? *Stoch. Sch.* 30(2), 13–20.
- Henze, N. (2018). *Stochastik für Einsteiger*. 12. Auflage, Springer Spektrum, Wiesbaden.
- Knapp, Th. R. (1985). Instances of Simpson's paradox. *The College Mathem. J.* 16(3), 209–211.

- Kocik, J. (2001). Proof without words: Simpson's paradox. *Mathem. Magazine* 74(5), 339.
- Künzel, E. (1991). Über Simpsons Paradoxon. *Stoch. Sch.* 11(1), 54–62.
- Lesser, L.M., und Pearl, D.K. (2020). Statistisches Edutainment: Die Umkehrung von Vergleichen. *Stoch. Sch.* 40(2), 30–34.
- Meyer, J. (1995). Einfache Paradoxien der beschreibenden Statistik. *Stoch. Sch.* 15(2), 27–50.
- Motzer, R. (2009). Ein kleines Simpson-Paradoxon bei den Ergebnissen von PISA-E. *Stoch. Sch.* 30(2), 22–25.
- Newson, G. (1991). Simpson's paradox revisited. *The Mathem. Gazette* 75(473), 290–293.
- Parlides, M.G., und Perlman, M.D. (2009). How likely is Simpson's paradox? *The Amer. Statistician* 63(3), 226–233.
- Reintjes, R., de Boer, A., van Pelt, N., und Mintjes-de Groot, J. (2000). Simpson's paradox: An example from Hospital Epidemiology. *Epidemiology* 11(1), 81–83.
- Samuels, M.L. (1993). Simpson's paradox and related phenomena. *J. Amer. Statist. Ass.* 88(421), 81–88.
- Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. B* 13(2), 238–241.
- von Kügelgen, J., Greseke, L., und Schölkopf, B. (2020). Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects. *arXiv:2005.07180v2* [stat.AP]
- Wagner, C.H. (1982). Simpson's paradox in real life. *The Amer. Statistician* 36(1), 46–48.
- Wardrop, E.L. (1995). Simpson's paradox and the Hot Hand in Basketball. *The Amer. Statistician* 49(1), 24–28.

Anschrift des Verfassers:

Prof. i.R. Dr. Norbert Henze  
 KIT Distinguished Senior Fellow  
 Institut für Stochastik  
 Karlsruher Institut für Technologie (KIT)  
 Englerstr. 2  
 76131 Karlsruhe  
 henze@kit.edu