# Koalitionsanalyse KOALA: Ein Werkzeug zur Berechnung von Wahrscheinlichkeiten einer Sitze-Mehrheit potentieller Regierungskoalitionen auf Basis aktueller Umfrageergebnisse

M.Sc. Andreas Bender: andreas.bender@stat.uni-muenchen.de
M.Sc. Alexander Bauer , Dr. André Klima, Prof. Helmut Küchenhoff

Statistisches Beratungslabor
Institut für Statistik LMU München

31. Juli 2017

**Zusammenfassung**

Wir stellen ein Verfahren vor, mit dem aus aktuellen Umfrageergebnissen Wahrscheinlichkeiten für Mehrheiten bestimmter Koalitionen abzuleiten sind. Unser Verfahren berücksichtigt die Unsicherheit, die mit der Tatsache verbunden ist, dass es sich bei den Umfragen um Stichproben handelt. Mögliche Abweichungen zwischen Umfrageergebnissen und dem tatsächlichen Wahlverhalten sind sehr schwer zu quantifizieren und werden hier nicht berücksichtigt. Es handelt sich also jeweils um Wahrscheinlichkeiten, die aus dem aktuellen Stimmungsbild abgeleitet werden. Diese werden mithilfe eines Bayesianischen Ansatzes unter der Annahme berechnet, dass es sich bei den Umfragen näherungsweise um einfache Zufallsstichproben handelt. Weiterhin wird eine Zusammenfassung mehrerer Umfragen vorgeschlagen, wobei Abhängigkeiten zwischen den Umfragen berücksichtigt werden.

# 1 Einführung

Bei aktuellen Umfrageergebnissen werden in der Regel von den Instituten für die einzelnen Parteien Prozentwerte angegeben, die eine Schätzung der entsprechenden Anteile darstellen. Zusätzlich wird in der Regel der Stichprobenumfang angegeben und – häufig im Kleingedruckten – Angaben zur Genauigkeit der Schätzung gemacht. Diese Umfragen spiegeln aktuelle Stimmungen und Absichten wider und das Wahlergebnis kann davon durchaus erheblich abweichen.

Daher wurde in der Politikwissenschaft viel über den Zusammenhang zwischen Stimmungen und Umfragen auf der einen Seite und den tatsächlichen Wahlergebnissen auf der anderen Seite geforscht. Im deutschen System sind insbesondere Koalitionspräferenzen von zentraler Bedeutung, da sie bestimmte Formen der strategischen Stimmvergabe steuern (siehe z.B. Pappi and Thurner (2002), Shikano et al. (2009)).

Weiter werden Strategien zur Wahlprognose diskutiert, die Daten zu früheren Wahlen nutzen und damit den Zusammenhang zwischen Unfrageergebnissen und dem tatsächlichen Wahlausgang mit Hilfe von statistischen Modellen analysieren. Zum Teil nutzen Umfrageinstitute solche Daten, um aus ihren Umfrageergebnissen Prognosen zum Wahlausgang durchzuführen. Die Forschungsgruppe Wahlen spricht in diesem Zusammenhang von *Projektion*, bei der solche Aspekte berücksichtigt werden. Die reinen Umfrageergebnisse werden als politische *Stimmung* bezeichnet.

In einem anderen Ansatz wird versucht mehrere Prognosen, die mit alternativen Methoden gewonnen worden sind, zu einer Gesamtprognose zu kombinieren (siehe PollyVote). Die angesprochenen alternativen Ansätze umfassen dabei klassische Wahlumfragen, Prognose- bzw. Wettmärkte, Expertenbefragungen sowie quantitative Modelle.

Ein Ansatz aus der Regressionsanalyse versucht den Wahlausgang anhand verschiedener Einflussgrößen, etwa der Langzeit-Unterstützung einer Partei oder der Dauer, die eine Regierungspartei an der Macht ist, zu prognostizieren (siehe z.B. Norpoth and Gschwend (2010)). Thomas Gschwend und sein Team veröffentlichen zur Bundestagswahl 2017 eigene Berechnungen, einschließlich der Wahrscheinlichkeit bestimmter Koalitionen (siehe zweitstimme.org).

Wir stellen im Folgenden eine Strategie vor, die Umfragergebisse weiterführend zu analysieren. Dabei interessiert uns die zentrale Frage, wie groß die Wahrscheinlichkeiten für bestimmte Mehrheitsverhältnisse sind. Diese bezieht sich auf das aktuelle Stimmungsbild und berücksichtigt nicht mögliche Veränderungen bis zur Wahl.

# 2 Vorgehen

Die aktuellen Umfrageergebnisse weisen eine erhebliche Unsicherheit auf, da jeweils nur ein geringer Teil der Wähler befragt wird. So gibt die Forschungsgruppe Wahlen bei ihren Umfragen einen Schwankungsbereich von 3% bei großen Parteien und einen Schwankungsbereich von 2% bei kleinen Parteien an. Daraus folgt auch, dass das Zustandekommen von bestimmten Mehrheiten unsicher ist. Wenn z.B. der Anteil der FDP in der Nähe von 5% geschätzt wird, hängt die Frage des Zustandekommens einer Mehrheit für Schwarz-Gelb auch sehr stark von der Frage ab, ob die FDP in den Bundestag kommt. Wir wollen nun auf Grund der Umfrageergebnisse Wahrscheinlichkeiten für bestimmte Mehrheitsverhältnisse berechnen.

Dabei haben wir uns für ein Vorgehen nach dem Prinzip der Bayesianischen Statistik (siehe z.B. Held (2008)) entschieden, welches sich für die vorliegende Fragestellung besonders gut eignet. Hierbei wird die Unsicherheit über die (unbekannten) Anteile der Parteien durch eine Wahrscheinlichkeitsverteilung modelliert. Diese Wahrscheinlichkeitsverteilung erhält man aus den jeweiligen Umfragedaten bzw. deren *Likelihood* sowie der *a priori* vorhandenen bzw. angenommenen Information über die Anteile – berücksichtigt in Form einer sog. *Priori-Verteilung*. Das Wissen über die Verteilung der Anteile nach Auswertung der Daten spiegelt sich dann in der *Posteriori-Verteilung* wider, einer Kombination aus Vorwissen und Beobachtung, wobei Vorwissen und Beobachtung je nach Fragestellung unterschiedlich stark ins Gewicht fallen können. Wir verwenden eine nichtinformative *Priori-Verteilung*, d. h. wir verzichten auf die Einbringung von Vorwissen.

Wir betrachten also die jeweils aktuellen Umfrageergebnisse als die vorhandene Information und bringen weder vorhergehende Umfragewerte noch die Ergebnisse vorangegangener Wahlen in die Berechnung der aktuellen Koalitionswahrscheinlichkeiten ein. Aus der *Posteriori-Verteilung* können wir dann, unter Berücksichtigung der Regeln für die Zusammensetzung des Bundestags, direkt Wahrscheinlichkeiten für das Zustandekommen bestimmter Mehrheiten berechnen. Weiter führen wir diese Berechnungen für eine Zusammenfassung der aktuellsten Umfragen der letzten 14 Tage durch („gepoolte Umfrage“), die im nächsten Abschnitt diskutiert wird. Weitere Details zur Berechnung sind in den Abschnitten 4 und 6 zu finden.

# 3 Zusammenfassung mehrerer Umfragen

Umfragen werden vor (Bundestags-)Wahlen von mehreren Instituten veröffentlicht. Neben der getrennten Betrachtung der einzelnen Umfragen kombinieren wir die vorliegenden Informationen zu einer sog. *gepoolten Umfrage*. Durch das Mehr an Information ist eine Verringerung der Schätzunsicherheit – und daher auch genauere Ergebnisse für die möglichen Mehrheiten von Koalitionen – zu erwarten. Daher werden die einzelnen Umfragen, gewichtet mit den von den Instituten angegebenen Stichprobenumfängen, zusammengefasst. Ginge man davon aus, dass die einzelnen Umfragen voneinander unabhängig sind, so würde das weitere Vorgehen dem bzgl. der Einzelumfragen entsprechen.

Weitere Untersuchungen zeigen jedoch, dass die einzelnen Umfragen nicht als unabhängig voneinander angenommen werden können. Daher wurde eine Korrektur vorgenommen, welche die Abhängigkeiten zwischen den Umfragen berücksichtigt. Details zu diesen Berechnungen finden sich in Abschnitt 5. Zu beachten ist, dass nur Umfragen zusammengefasst werden, die einen geringen zeitlichen Abstand haben. Nur dann bildet die Zusammenfassung die aktuelle Stimmung korrekt ab. Gibt es starke durch besondere Ereignisse erklärbare Veränderungen zwischen den einzelnen Umfragen, so sollten die Umfragen besser einzeln betrachtet werden.

# 4 Theoretischer Hintergrund

Um prädiktive Aussagen über die Wahrscheinlichkeiten verschiedener Koalitionen bedingt auf aktuelle Umfrage-Ergebnisse treffen zu können nutzen wir einen bayesianischen Ansatz. D.h. wir haben *a priori* Annahmen bzgl. der Verteilung der Anteile für die einzelnen Parteien $p(\theta)$ und Annahmen über die Datenverteilung $f(x|\theta)$, woraus sich nach dem Satz von Bayes die *Posteriori-Verteilung* für die Anteile der einzelnen Parteien ergibt:

$$f(\theta|x) = \frac{f(\theta, x)}{f(x)} = \frac{f(x|\theta)p(\theta)}{f(x)}$$

Zur Modellierung von Wahlergebnissen gehen wir von einer Mulitinomialverteilung der Ergebnisse einer Umfrage aus. Wir gehen von $n$ befragten Personen aus und betrachten in unserem Modell $k = 7$ verschiedenen Alternativen, *Union (CDU/CSU)*, *SPD*, *FDP*, *GRÜNE*, *DIE LINKE*, *AfD* und *Sonstige*. Seien $X_j, 1 \leq j \leq k$ die Zufallsvariablen, welche die Anzahl der Stimmen pro Alternative angeben. Dann nehmen wir an, dass $(X_1, \ldots, X_k)$ einer Multinomialverteilung mit den Parametern $n$ und $(\theta_1, \ldots, \theta_k)$ genügt:

$$\boldsymbol{X_1, \ldots, X_k} \sim Multinomial(n, \theta_1, \ldots, \theta_k)$$

Die Parameter $\theta_j$ sind dann die Wahrscheinlichkeiten, die Partei $j$ anzugeben. Geht man von einer einfachen Zufallsstichprobe bei einer großen Grundgesamtheit aus, so entsprechen diese Wahrscheinlichkeiten $\theta_j$ genau den jeweiligen Anteilen von Wählern der Partei $j$ in der Grundgesamtheit, siehe dazu z.B. Kauermann and Küchenhoff (2011).

Nun wählen wir die zur Multinomialverteilung konjugierte Verteilung als *Priori-Verteilung*. Dies ist in diesem Fall die Dirichletverteilung. *A priori* nehmen wir also an, dass der Parameter(-vektor) einer Dirichletverteilung folgt

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$$

Dabei geben die $\alpha_1, \ldots, \alpha_k$ das Vorwissen über die $\theta_j$ an. Wir wählen eine nicht informative *Priori-Verteilung* (Jeffreys prior) mit $\alpha_1 = \ldots = \alpha_k = 1/2$. Für die *Posteriori-Verteilung* ergibt sich in diesem Fall

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{x}) &\propto \mathrm{L}(\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \\ &\propto \prod_{j=1}^{k} \theta_j^{x_j} \cdot \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \\ &\propto \prod_{j=1}^{k} \theta_j^{x_j + \alpha_j - 1} \end{aligned}$$

d.h. die *Posteriori-Verteilung* ist wiederum eine Dirichletverteilung mit den Parametern

$$x_1 + 1/2, \ldots, x_k + 1/2$$

Diese Posteriori-Verteilung enthält nach dem Prinzip der Bayes-Statistik das gesamte Wissen über den Parametervektor $\boldsymbol{\theta}$. Daher können aus dieser Verteilung Wahrscheinlichkeitsaussagen über die Parameter abgeleitet werden. Uns interessiert z.B. die Wahrscheinlichkeit dafür, dass die Parameterkonstellation (d.h. die Wähleranteile in der Grundgesamtheit) zu einer Mehrheit von CDU/CSU-FDP führt. Da dies nicht direkt berechenbar ist, verwenden wir sog. Monte-Carlo-Simulationen, um die Wahrscheinlichkeit zu bestimmen.
Pro Simulationsschritt ziehen wir Zufallszahlen aus der *Posteriori-Verteilung*. Diese stehen dann für die jeweiligen Anteile der oben genannten Parteien. Daraus bestimmen wir nach dem Verfahren von Sainte-Laguë/Schepers die Verteilung der Sitze im Bundestag, mit welcher sich wiederum die Wahrscheinlichkeiten für Mehrheiten bestimmter Koalitionen berechnen lassen. Mögliche Überhangs- und Ausgleichsmandate bleiben hier unberücksichtigt, sollten nach dem neuen Wahlrecht aber kaum eine Rolle bei der Bestimmung der Mehrheitsverhältnisse spielen.

# 5 Details zur Zusammenfassung von mehreren Umfragen

Die Unsicherheit einer einzelnen Umfrage lässt sich unter der Annahme, dass es sich um eine einfache Zufallsstichprobe handelt, über eine Multinomialverteilung bestimmen (siehe oben): Bei einer einzelnen Umfrage $i$ wird angenommen, dass die Anteile der Parteien $(X_1, \ldots, X_k)$ einer Multinomialverteilung mit den Parametern $n_i$ und $(\theta_1, \ldots, \theta_k)$ genügen:

$$\boldsymbol{X_1, \ldots, X_k} \sim Multinomial(n_i, \theta_1, \ldots, \theta_k)$$

Dabei sind die $n_i$ die jeweiligen Stichprobenumfänge und die Parameter $(\theta_1, \ldots, \theta_k)$ die unbekannten Anteile der Parteien in der Grundgesamtheit. Wir gehen davon aus, dass sich diese Anteile zwischen zwei Umfragen nicht wesentlich ändern. Bei der Zusammenfassung von $I$ Umfragen mehrerer Institute zu einer gepoolten Umfrage gilt für die Verteilung unter Annahme der Unabhängigkeit:

$$\boldsymbol{X_1, \ldots, X_k} \sim Multinomial(\sum_{i=1}^{I} n_i, \theta_1, \ldots, \theta_k)$$

Die Berechnung der Wahrscheinlichkeiten kann nun einfach mit dem neuen Stichprobenumfang $\sum_{i=1}^{I} n_i$ durchgeführt werden. Es zeigt sich allerdings, dass die Annahme der Unabhängigkeit der Umfragen problematisch ist. Wir bestimmen daher eine Schätzung für die Korrelation zwischen den Umfragen. Dazu betrachten wir nun die Differenz zwischen zwei Umfragen. Allgemein gilt für 2 Zufallsgrößen $X_1$ und $X_2$:

$$\begin{aligned} Var(X_1 - X_2) &= Var(X_1) + Var(X_2) - 2 \cdot Cov(X_1, X_2) \\ \Rightarrow \quad Cov(X_1, X_2) &= 0.5 \cdot (Var(X_1) + Var(X_2) - Var(X_1 - X_2)) \end{aligned}$$

Definiert man beispielsweise $X_1$ als den Anteil einer bestimmten Partei in der ersten Umfrage und $X_2$ als den Anteil der gleichen Partei in Umfrage 2, so können deren Varianzen über die Binomialverteilung geschätzt werden. Die Varianz der Differenz $X_1 - X_2$ kann aus den beobachten Differenzen zwischen den geschätzten Umfrageergebnissen geschätzt werden. Aus der Kovarianz lässt sich nun direkt die Korrelation $corr(X_1, X_2)$ zwischen den beiden Zufallsvariablen (Umfragen) berechnen. Zugleich lässt sich über den gleichen Zusammenhang über die Korrelation – für zwei spezifische Umfragen unter Annahme einer gleichbleibenden Korrelation – die Kovarianz zwischen den beiden Umfragen berechnen. Um das Vorgehen für die Bayesianische Schätzung möglichst einfach zu gestalten, verwen-

den wir das Konzept des effektiven Stichprobenumfangs. Da die Varianz der Binomialverteilung direkt proportional zum Stichprobenumfang $n$ ist, definiert man das Verhältnis der Varianz für eine Stichprobe vom Umfang 1 und der mit Hilfe obiger Überlegungen berechneten Varianz für die gepoolte Stichprobe. Es gilt:

$$n_{eff} = \frac{\text{Varianz(gepoolte Stichprobe)}}{\text{Varianz(Stichprobe mit n=1)}}$$

Mit dieser Berechnung wird der effektive Stichprobenumfang so gewählt, dass die Varianz der Binomialverteilung der gepoolten Stichprobe, welche einen effektiven Stichprobenumfang von $n_{eff}$ hat, der Varianz der gewichteten Summe der Einzelumfragen unter Berücksichtigung der geschätzten Kovarianz entspricht. Dieser effektive Stichprobenumfang wird dann zur Berechnung der *Posteriori-Verteilung* der Parameter und zur Schätzung der Wahrscheinlichkeiten genutzt.

In der praktischen Umsetzung des dargestellten Vorgehens ergibt sich das Problem, dass für eine sinnvolle Berechnung der Korrelation zwischen zwei Umfragen, diese in etwa den gleichen Befragungszeitraum abdecken müssen. Dies ist jedoch vor allem im Zeitraum sechs Monate vor der Wahl nicht für alle Institute der Fall. Weiterhin ist für eine sinnvolle Schätzung der Varianz der Differenz eine entsprechende Zahl an vergleichbaren Umfragen für die Differenzenbildung notwendig. Als besonders vielversprechend erwiesen sich dabei die Umfragen von Emnid und Forsa, welche relativ regelmäßig veröffentlicht werden und welche auch oft Überschneidungen im Befragungszeitraum aufweisen.

Anhand dieser Daten wurde letztendlich eine Korrelation in der Größenordnung von $\text{corr}(\text{Emnid}, \text{Forsa}) \approx 0.7$ für die Union geschätzt. Dabei wurden jeweils 20 Umfragen der beiden Institute miteinander verglichen, für die Berechnung der theoretischen Varianzen der beiden Institute wurde jeweils der mittlere Umfragewert und der mittlere Stichprobenumfang im betrachteten Zeitraum genutzt. Eine weiterführende Betrachtung zeigt, dass die Korrelation in dieser Größenordnung nicht für alle Parteien und Zeiträume gilt. Jedoch kann oft von einer mittleren positiven Korrelation ausgegangen werden. Aus diesem Grund nehmen wir bei den Berechnungen des effektiven Stichprobenumfangs pauschal den Wert $\text{corr}(\text{Institut } 1, \text{Institut } 2) = 0.5$ an. Die Berechnung des effektiven Stichprobenumfangs wird anhand der Unionsergebnisse durchgeführt. Exemplarisch durchgeführte Berechnungen für andere Parteien zeigten, dass es zwischen den parteispezifischen effektiven Stichprobenumfängen nur geringe Unterschiede gibt.

# 6 Ergänzungen

## 6.1 Algorithmus zur Sitzeverteilung

Zur Bestimmung der Sitzeverteilung im Parlament gegeben die aktuellen Umfragewerte bzw. gegeben die simulierten Anteile wird das Verfahren von Sainte-Laguë/Schepers verwendet. Es gibt hierbei verschiedene Berechnungsverfahren; hier wird das sog. Rangmaßzahlenverfahren verwendet.

Dabei wird in folgenden Schritten vorgegangen:

1. Bilde die *Rangmaßzahlen* $R_{ik} = (i - 0.5) \cdot V_{tot}/V_k$ mit $i = 1, 2, 3, \ldots$, $V_{tot}$ die Gesamtzahl abgegebener Stimmen und $V_k$ die Anzahl Stimmen für Partei $k \in \{\text{CDU/CSU}, \text{SPD}, \ldots\}$ für alle Parteien, welche die 5% Hürde überschreiten.

2. Bestimme die Ränge der Rangmaßzahlen $R_{ik}$ und verteile diese bis alle (598) Sitze vergeben sind. Im Falle von Bindungen wird der Rang in unserer Simulation zufällig vergeben.

## 6.2 Algorithmus zur Monte–Carlo–Simulation

Bevor der in Abschnitt 6.1 beschriebene Algorithmus zur Verteilung der Sitze angewendet werden kann, müssen zunächst die Anteile, welche die verschiedenen Parteien erreichen, simuliert werden. D.h. wir müssen Zufallszahlen aus der Dirichlet-Verteilung ziehen. Die an den Algorithmus übergebenen Parameter (Konzentrations-Gewichte) ergeben sich dabei, wie in Abschnitt 4 beschrieben, aus der *Posteriori-Verteilung*.

Die Konzentrations-Gewichte für das Ziehen von Zufallszahlen ergeben sich dann aus den Stimmen, die jede Partei in der Umfrage bekommen hat sowie den Priori-Parametern, die aufgrund der Verwendung von Jeffrey's Priori auf 1/2 gesetzt worden sind. Damit ziehen wir Zufallszahlen aus einer Dirichlet-Verteilung. Da von den Instituten üblicherweise nur die gerundeten Anteile zur Verfügung gestellt werden, addieren wir auf jeden Anteil zusätzlich noch eine gleichverteilte Zufallsvariable $r_\gamma \sim G[-\gamma, \gamma]$, um die Rundungsfehler zu berücksichtigen. Die Größe $\gamma$ wird dabei auf den Wert 0.5% gesetzt. Für jede dieser Zufallsziehungen kann nun wie in Abschnitt 6.1 dargestellt die Sitzeverteilung im Parlament berechnet, und daraus wiederum die Wahrscheinlichkeiten für bestimmte Koalitionen bestimmt werden.

Die Umsetzung des Verfahrens erfolgte in der Programmiersprache **R** (R Core Team, 2017). Die Open Source Implementierung wird im R-Paket **coalitions** zur Verfügung ge-

stellt und kann unter folgender Adresse eingesehen werden: https://adibender.github.io/coalitions/

# Literatur

Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes*. Heidelberg: Spektrum Akademischer Verlag.

Kauermann, G. and H. Küchenhoff (2011). *Stichproben: Methoden und praktische Umsetzung mit R*. Methoden und praktische Umsetzung mit R. Heidelberg [u.a.]: Springer.

Norpoth, H. and T. Gschwend (2010). The chancellor model: Forecasting German elections. *Special Section: European Election Forecasting 26*(1), 42–53.

Pappi, F. U. and P. W. Thurner (2002). Electoral behaviour in a two-vote system: Incentives for ticket splitting in German Bundestag elections. *European Journal of Political Research* (41), 207–232.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shikano, S., M. Herrmann, and P. W. Thurner (2009). Strategic Voting under Proportional Representation: Threshold Insurance in German Elections. *West European Politics 32*(3), 634–656.