Statistik für alle

Walter Krämer

Die 101 wichtigsten Begriffe anschaulich erklärt

SACHBUCH

Springer Spektrum

Statistik für alle

Walter Krämer

Statistik für alle

Die 101 wichtigsten Begriffe anschaulich erklärt

Springer Spektrum

Prof. Dr. Walter Krämer
Institut für Wirtschaftschafts-
und Sozialstatistik
Technische Universität Dortmund
Dortmund, Deutschland

ISBN 978-3-662-45030-7 ISBN 978-3-662-45031-4 (eBook)
DOI 10.1007/978-3-662-45031-4

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer Spektrum
Das Buch basiert auf dem Titel „Statistik für die Westentasche", der 2002 bei Piper Verlag GmbH, München erschien.

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Springer-Verlag GmbH Berlin Heidelberg ist Teil der Fachverlagsgruppe Springer Science+ Business Media
(www.springer.com)

Vorwort

Mit der Statistik ist es wie mit der städtischen Müllabfuhr: Ihre Bedeutung für unser Sozialwesen ist umgekehrt proportional zu der Anerkennung, die sie in den Medien und in der internen Wertehierarchie der meisten Menschen erfährt. Daran habe ich mich gewöhnt. Ich bin seit 30 Jahren Professor für Statistik und schalte automatisch ab, wenn es wieder im Sportfernsehen heißt: „Wie oft hat eigentlich Bayern München gegen den BVB verloren? Fragen wir mal unseren Statistiker!" So als wäre das etwas für Leute, die gerade mal Eins und Eins zusammenzählen, aber sonst nicht viel zustande bringen können.

In Wahrheit ist natürlich Statistik das faszinierendste Thema, das es für einen an Wahrheit und nicht an Wünschen interessierten Menschen gibt. Um hier auf intelligente Weise mitzureden, braucht man weder das Abitur noch eine spezielle mathematische Begabung, der gute Wille reicht. Allerdings ist es nützlich, die wichtigsten Begriffe zu kennen, um nicht immer wieder auf die Schwulstrhetorik der Datenmanipulateure hereinzufallen. Speziell der Signifikanztest-Unfug wird immer wieder gern genutzt, um blauäugige Zeitgenossen hinters Licht zu führen. Als Gegenmittel gibt es dieses Taschenwörterbuch. Es ist eine im Umfang mehr als verdoppelte, aktualisierte und nochmals

auf hoffentlich leserfreundliche Art und Weise umgeschriebene Fassung meines kleinen Lexikons „Statistik für die Westentasche“, das vor 15 Jahren im Piper-Verlag in München erschienen ist. Insbesondere habe ich allen unnötigen Fachjargon entfernt und auch Themen aufgenommen, die mir seinerzeit noch fremd gewesen, aber inzwischen ins Zentrum der öffentlichen Aufmerksamkeit gewandert sind. Wer hätte etwa damals ahnen können, wie wichtig es heute für ganze Länder ist, von der Rating-Agentur Moody's ein Aaa zu erhalten, oder dass man inzwischen keine Kreditkarte mehr bekommt, ohne vorher Objekt einer Scorekarten-Evaluation zu sein? Daran sehen wir auch schon: Statistik betrifft die große Politik und die kleinen Sparer und Konsumenten gleichermaßen. Und in dem Umfang, wie wir alle unsere digitalen Fingerabdrücke im weltweiten Netz hinterlassen, werden auch die unter uns, die nicht an Gott glauben, zumindest als statistische Einheiten bis zum Ende aller Tage weiterleben. Allein schon deshalb ist es nützlich, wenn man weiß, um was es dabei geht.

Dortmund, im Herbst 2014 Walter Krämer

Danksagung

Wie bei allen meinen Statistikbüchern waren mir auch diesmal meine Mitarbeiter an der Fakultät Statistik der TU Dortmund eine große Hilfe. Kira Ahlhorn und Etienne Theising haben mir bei der Datenrecherche und bei der Herstellung von Schaubildern viel Arbeit abgenommen, Carmen van Meegen, Robert Löser und Simon Neumärker haben ebenfalls viele Tabellen und Quellen recherchiert, Eva Brune hat zahlreiche Verbesserungen von Verständlichkeit und Stil sowie eigene Textvorschläge eingebracht, Maarten van Kampen, Matthias Arnold, Marianthi Neblik und Katharina Pape haben beim Korrekturlesen geholfen, und Sebastian Voß hat einige eher mathematische Passagen kontrolliert (und dabei auch den einen oder anderen Fehler aufgedeckt). Auch Clemens Heine vom Springer-Verlag hat durch seine Korrekturvorschläge die Lesbarkeit verbessert. Und als weitere Korrekturleser und Rechercheure waren wie bei allen meinen Büchern auch diesmal wieder Denis und Eva Krämer unterwegs. Ich danke allen Helfern für die großzügige Unterstützung schließe mit der durchaus ernstgemeinten Standardfloskel, dass verbleibende Fehler und Unklarheiten allein dem Autor anzulasten sind.

Inhaltsverzeichnis

Statistik für alle 1

Achsenmanipulation 1
Adäquationsproblem 5
Aktienkurse 6
Äquivalenzskala 9
Arbeitslosenquote 11
Arithmetisches Mittel 13
Armutsmaße 16
Ausfallratings 20
Ausreißer 22
Balkendiagramm 24
Bayes-Statistik 25
Bedingte Wahrscheinlichkeiten 27
Benford-Gesetz 31
Bereinigtes Lohndifferential 34
Big Data 35
Biometrischer Fingerabdruck 38
Binomialverteilung 40
Bruttosozialprodukt 41
Chaos 44
Chartanalyse 47
Chartjunk 49
Datenschutz 51
DAX 52
Dow-Jones 55
Epidemiologie 57

Erwartungswert 59
Exponentielles Glätten 61
Fehlende Werte 63
Fehler 1. Art 65
Fragebögen 67
Fruchtbarkeitsziffer 70
Geldmenge 73
Geometrisches Mittel 75
Gewichtete Mittelwerte 77
Gesetz der Großen Zahl 79
Gleitende Durchschnitte 82
Harmonisches Mittel 83
Histogramm 86
Indirekte Befragung 88
Innumeratentum 91
Intervallskala 93
Itemanalyse 94
Kartogramme 95
Kaufkraftparitäten 97
Klinische Studien 99
Klumpenstichprobe 101
Konfidenzintervalle 103
Konkurrierende Risiken 104
Kontrollkarten 108
Korrelationskoeffizient 109
Kreuztabelle 113
Kurvendiagramme 115
Lebenserwartung 118
Logarithmische Skala 119
Logistische Regression 122
Lohnquote 123
Lorenzkurve 124
Median 127
Meinungsumfragen 129
Mengenindices 133
Methode der Kleinsten Quadrate 135

Mikrozensus 137
Multiplikationsregel 139
Nonsenskorrelation 141
Normalverteilung 144
Optionsbewertung 145
Paneldaten 148
Polizeiliche Kriminalstatistik 149
Preisindices 150
Quantile 153
Random Walk 154
Regression zum Mittelwert 156
Regressionsanalyse 159
Saisonbereinigung 163
Scheinpräzision 165
Scorekarten 167
Signifikanztests 168
Simpson-Paradox 171
Standardabweichung 173
Stichproben 176
Streudiagramm 178
Terms of Trade 180
Tortendiagramm 181
Trend 183
Trendextrapolation 185
t-Test 186
Varianzanalyse 187
Verlaufsdaten 190
Volkseinkommen 191
Volkszählung 192
Wahrscheinlichkeitsprognosen 195
Warenkorb 197
Wechselwirkungen 199
Weißes Rauschen 200
Wohlfahrtsmaße 202
Zahlungsbilanz 203
Zensierte Daten 206

Zeitreihen 207
Zipfsches Gesetz 209
Zufall 212
Zufallszahlen 217

Sachverzeichnis 219

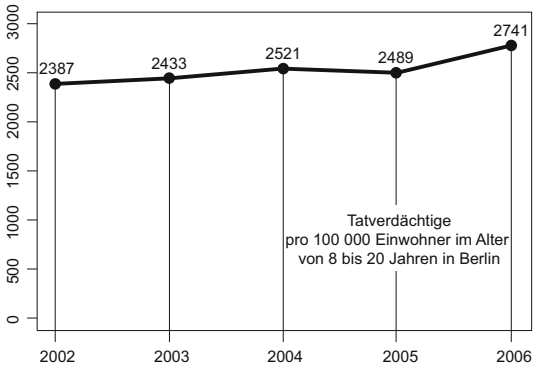
Statistik für alle

Achsenmanipulation

Fast alle Datengrafiken haben mindestens eine Achse. Darauf sind die möglichen Werte der jeweils interessierenden Variablen abgetragen. Ein Beispiel ist die folgende Grafik der Kriminalitätsentwicklung in Berlin. Hier ist auf der – hier nur gedachten – senkrechten Achse die Anzahl der jährlich erfassten Delikte pro 100.000 potentielle Täter abgetragen. Und wie wir sehen, nimmt die sehr stark zu.

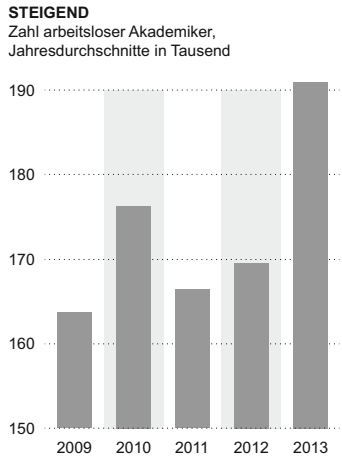
W. Krämer, *Statistik für alle*, DOI 10.1007/978-3-662-45031-4_1,

In Wahrheit nehmen die Delikte aber nur sehr wenig zu. Der falsche Eindruck eines starken Anstiegs ist eine optische Täuschung, hervorgerufen durch das Abschneiden der senkrechten Achse bei 2200. Würde die bei Null beginnen, wie es sich gehört, käme eine Grafik wie die folgende heraus:

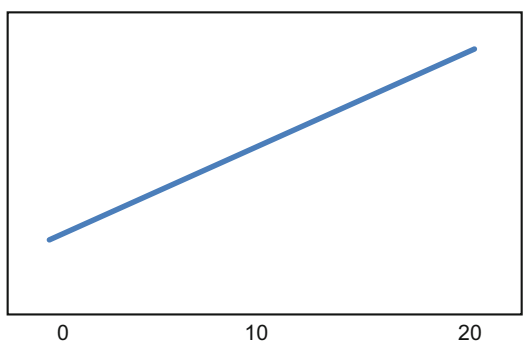


Quelle: Polizeiliche Kriminalstatistik Berlin 2006; Rohheitsdelikte

Auch bei Säulendiagrammen ist dieses Abschneiden der senkrechten Achse beliebt. Die folgende Grafik suggeriert starke Schwankungen, verbunden mit einem steilen Anstieg, der deutschen Akademikerarbeitslosigkeit. In Wahrheit sind sowohl die Schwankungen wie der Anstieg minimal. Auch hier kommt der falsche erste Eindruck nur so zustande, dass die langen Beine der Säulen nicht zu sehen sind.

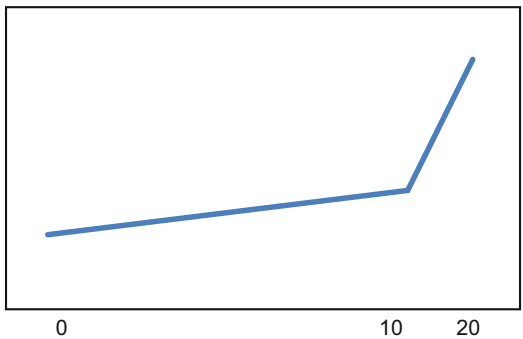


Fortgeschrittene Achsenmanipulateure verbiegen auch die waagerechte Achse. Angenommen, eine Variable entwickelt sich über die Zeit wie folgt:

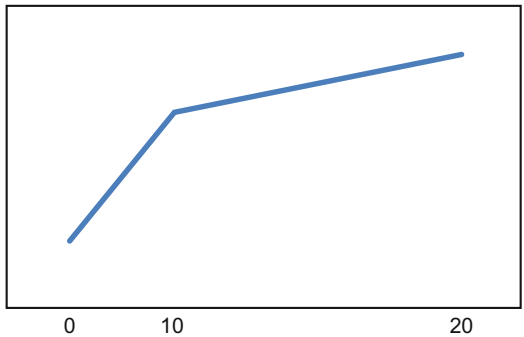


Jede Periode geht es gleichmäßig bergauf. Angenommen, die Variable steht für etwas Angenehmes, etwa das Volkseinkommen, und in Periode 10 wechselt die Regierung. Vorher die anderen, danach wir. Wie stellen wir diese Entwicklung in ein für uns günstigeres Licht?

Nichts einfacher als das: der Achsenabschnitt von 10 bis 20 wird gestaucht, der von 0 bis 10 gedehnt:



Oder die Variable ist etwas, das man ungern wachsen sieht, etwa das Preisniveau. Auch da stellen wir uns in ein günstigeres Licht. Jetzt wird der Achsenabschnitt von 0 bis 10 gestaucht und der von 10 bis 20 gedehnt. Dergleichen Achsenmanipulationen erfordern schon eine gewisse kriminelle Energie und kommen deshalb eher selten vor.



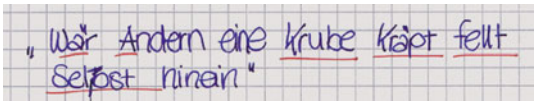
Adäquationsproblem

Mit Adäquationsproblem ist gemeint, dass es in vielen Anwendungen der Statistik alles andere als offensichtlich ist, wie zentrale Begriffe wie Armut, Krankheit oder Kriminalität zu messen sind. Je nachdem, wie man diese Tatbestände definiert, sind mal mehr und mal weniger Menschen arm, kriminell, krank oder arbeitslos. Fast alle diese Themen kommen in eigenen Stichwortartikeln später nochmals vor.

Sehen wir uns hier einmal an, wie viele Analphabeten es in Deutschland gibt. Laut der Hamburger *Zeit* sind es 7,5 Millionen („7,5 Millionen Deutsche sind Analphabeten", Zeit Online vom 2. März 2011). Liest man aber weiter, so wird deutlich, dass dies sogenannte „funktionale Analphabeten" sind, also Menschen, die zwar einzelne Sätze, aber keine zusammenhängenden Texte lesen und schreiben können. Also nicht das, was man normalerweise unter Analphabeten versteht. Und nochmal etwas anders

ist die Legasthenie, d. h. eine möglicherweise genetisch verursachte Leser und Rechtschreibschwäche, wie sie selbst bei sonst hochintelligenten Menschen vorkommt.

Nicht notwendig ein Zeichen mangelnder Intelligenz



Zu Kaiser Wilhelms Zeiten machte man sich das Leben einfacher; ein Rekrut (bei Frauen hat man sich derartige Statistiken erspart) galt als Analphabet, wenn er bei der Unterschrift ein Kreuz statt Namen hinterließ. So gesehen ist es also überhaupt kein Wunder, dass die Zahl der Analphabeten seit dieser Zeit in Deutschland zugenommen hat.

Aktienkurse

Aktienkurse schwanken von Tag zu Tag, von Minute zu Minute, von Sekunde zu Sekunde. Den einen raubt das den Schlaf, den anderen füllt es den Geldbeutel.

Warum und wie die Aktienkurse schwanken, ist Gegenstand von heftigen Debatten. „Weil die Börsianer nicht wissen, was sie wollen", sagen die einen. „Weil sie nur zu gut wissen, was sie wollen," sagen die anderen.

Die anderen haben Recht. Ein richtiger, „gerechter" Aktienkurs ist der Barwert aller künftigen Erträge. Barwert heißt: künftige Erträge werden abgezinst. Wenn ich heute in einem Jahr 100 € habe, so ist mir das heute nur – sagen

wir – 98 € wert. Diese Erträge – vor allem Dividenden, aber auch Bezugsrechte und andere Ansprüche, die Aktionäre an ihre Gesellschaft haben – sind heute allenfalls in Ansätzen bekannt. Wer heute eine VW-, BASF- oder Daimler-Aktie besitzt, kann ohne allzu großes Zittern darauf hoffen, im nächsten Jahr rund einen Euro Dividende zu erhalten. Auch noch im übernächsten Jahr. Aber dann wird die Sache zusehends riskanter. Wer weiß, wie die Nachfrage nach den Produkten unserer Firma in 5 oder 6 Jahren aussieht? Und was in 30 oder 40 Jahren geschieht, das weiß der Liebe Gott allein.

Deshalb kann man nicht mit sicheren Erträgen rechnen. Stattdessen nimmt man die *erwarteten* Erträge. In einem effizienten Kapitalmarkt fließen in diese Erwartungen alle Informationen ein, die es aktuell zu einer Firma gibt: Ölpreise, Dollarkurs, Diskontsatz usw. – alles, was den „gerechten" Wert einer Aktie berühren könnte, ist schon im aktuellen Preis enthalten.

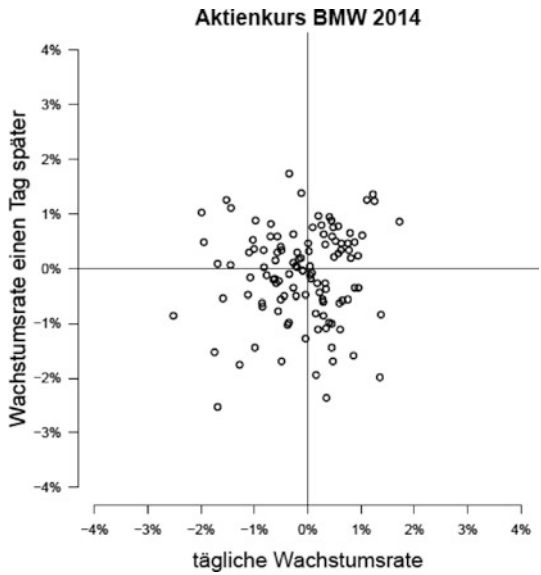
Falls nicht, könnten kluge Leute durch das Ausnützen von Unter- oder Überbewertungen schnell zu Milliardären werden. Und an der Börse gibt es viele kluge Leute. Indem diese bei einer Unterbewertung kaufen, treiben sie den Kurs nach oben. Indem diese bei einer Überbewertung verkaufen, treiben sie den Kurs nach unten. Damit sorgen sie dafür, dass mögliche Fehlbewertungen schnell verschwinden.

„Korrekte" Aktienkurse können sich damit nur ändern, wenn etwas Unerwartetes geschieht. Unerwartete Ereignisse haben aber die Eigenschaft, recht zufällig und chaotisch aufzutreten – sonst wären sie nicht unerwartet. Damit sind aber auch die Änderungen in den Aktienkursen unerwartet, zufällig und chaotisch. Nicht, weil an der Börse wild gewür-

felt würde. Sondern weil das wahre Leben würfelt, weil nur aktuell noch nicht bekannte Dinge die aktuellen Kurse ändern können.

Aktienkurse folgen also einen sogenannten „Random Walk“: Das ist eine Folge von Zufallszahlen, bei der man – grob gesprochen – nicht weiß, ob es beim nächsten Mal nach oben oder nach unten geht. Und das ist in einem korrekt bewerteten Aktienmarkt der Fall. Vom aktuellen Wert geht es nach oben oder nach unten, beides mit Wahrscheinlichkeit 1/2. Welcher dieser Fälle eintritt, weiß man heute nicht (denn wenn man es wüsste, wäre der aktuelle Kurs schon angepasst). Wenn es also heute nach oben geht, kann es morgen mit der gleichen Wahrscheinlichkeit weiter nach oben, aber auch nach unten gehen.

Die folgende Grafik zeigt dieses Verhalten am Beispiel des Aktienkurses von BMW. Abgetragen sind die täglichen relativen Kursänderungen für die ersten 50 Börsentage des Jahres 2014. Wie man sieht, folgt auf einen positiven Börsentag (das sind alle Punkte rechts von der Null) fast exakt genau so oft ein weiterer positiver wie ein negativer Börsentag.



Äquivalenzskala

Dieser seltsame technische Ausdruck hat eine große sozialpolitische Bedeutung. Wann immer uns im Herbst die regelmäßigen Horrormeldungen über Armut in Deutschland erschrecken, ist die Äquivalenzskala in der ersten Reihe mit dabei. Denn sie bestimmt im Wesentlichen mit, wie viel Geld eine Familie braucht, um nicht mehr arm zu sein.

Die seltsame Bestimmung der Armutsgrenze ist Gegenstand eines weiteren Stichwortartikels. Die nehmen wir hier einmal als gegeben hin. Der aktuelle Stand für Deutschland ist: Ein alleinstehender Erwachsener, gleich ob Mann

oder Frau, braucht pro Monat netto rund 1000 Euro, um nicht mehr arm zu sein (die konkreten Grenzen sind von Studie zu Studie leicht verschieden). Aber wie viel braucht ein Ehepaar? Oder eine alleinerziehende Mutter mit einem Kind? Oder ein Ehepaar mit drei Kindern? Auch ohne viel Statistik ist hier jedem klar: Ein Ehepaar braucht nicht das Doppelte, ein Ehepaar mit drei Kindern braucht nicht das fünffache einer Einzelperson. Schließlich braucht ein Haushalt mit zwei Personen keine zwei Waschmaschinen und keine zwei Kühlschränke, einer langt. Oder technisch ausgedrückt: Die Fixkosten verteilen sich bei größeren „Bedarfsgemeinschaften" auf mehr Köpfe (Das ist der Fachausdruck für Leute, die einen gemeinsamen Haushalt führen.) Auch der Bedarf an Wohnraum steigt nicht proportional, eine Küche und ein Badezimmer reichen weiterhin. Meistens jedenfalls. Aber wie viel mehr braucht nun eine Familie von vier tatsächlich?

Die Antwort liefert die Äquivalenzskala. Nach aktuellem Stand braucht jede weitere Person über 14 Jahre nochmals die Hälfte dessen, was die erste braucht. Und jedes Kind bis 14 geht mit nochmals weiteren 30 % in die Armutsgrenze der Bedarfsgemeinschaft ein. Dieses Schema ist auch als „Neue OECD-Skala" bekannt. Damit braucht ein Ehepaar mit zwei Kindern nicht $4 \cdot 1000$ Euro = 4000 Euro pro Monat, sondern nur

$$1000 + 500 + 300 + 300 = 2100 \text{ Euro}$$

pro Monat, um nicht mehr arm zu sein.

Die Bedeutung der Äquivalenzskala für die Armutsquote ist evident: Sind die nötigen Zusatzbeträge für weitere

Personen hoch, braucht man mehr Geld, um nicht mehr offiziell statistisch arm zu sein. Sind die nötigen Zusatzbeträge niedrig, reicht schon ein mäßiges Einkommen, um der Armut zu entkommen. In der alten QECD-Skala etwa wurde jeder weiteren Person über 14 Jahre 70 % der ersten, und jedem Kind 50 % als zusätzlichen Minimalbedarf zugestanden. Damit bräuchte unsere Familie mit zwei Kindern schon

$$1000 + 700 + 500 + 500 = 2700 \text{ Euro}$$

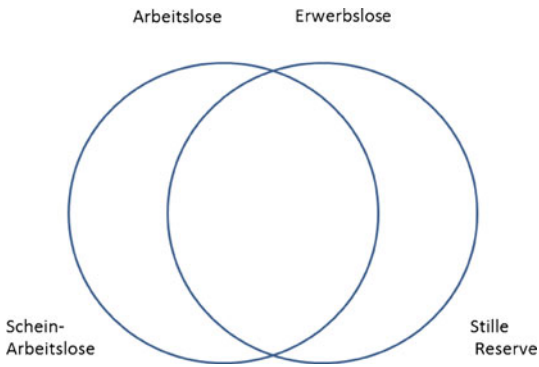
pro Monat, um nicht mehr arm zu sein. Damit wären nach dieser Sicht der Dinge weit mehr Menschen arm.

Arbeitslosenquote

Die bundesdeutsche Arbeitslosenstatistik ist ein viel beachteter und jeden Monat mit Hoffen, Bangen, zuweilen auch Resignation erwarteter Wirtschaftsindikator. Sie zählt als arbeitslos, wer (i) für mehr als 18 Stunden in der Woche und nicht nur vorübergehend Arbeit sucht, (ii) dem Arbeitsmarkt unmittelbar zur Verfügung steht, und (iii) als arbeitslos gemeldet ist. Außerdem muss die betreffende Person älter als 15 und jünger als 65 Jahre sein.

Das große Problem ist die Bedingung (iii). Denn viele als arbeitslos gemeldete suchen in Wahrheit gar keine Arbeit. Das wird jeder Arbeitgeber gerne bestätigen. Sie wollen die Zeit zwischen zwei Beschäftigungsverhältnissen überbrücken oder einfach nur Hartz IV kassieren. Und viele Menschen, die tatsächlich Arbeit suchen, sind nicht als arbeitslos gemeldet.

Diese Menschen heißen auch „stille Reserve“. Diese stille Reserve nimmt oft in einem Wirtschaftsabschwung zu (weil viele Arbeitssuchende die Hoffnung aufgeben und sich bei den Arbeitsämtern abmelden), in einem Wirtschaftsaufschwung aber ab: Jetzt fassen die Menschen wieder Mut und fragen Arbeit nach, mit dem Ergebnis, dass trotz Aufschwung die Arbeitslosenzahl nicht fällt, vielleicht sogar noch steigt.



Ein weiteres Problem ist die Bedingung (ii): dem Arbeitsmarkt unmittelbar zur Verfügung stehen. Danach sind Teilnehmer von Schul- und Umschulungsmaßnahmen, da nicht unmittelbar dem Arbeitsmarkt zur Verfügung stehend, offiziell nicht arbeitslos. Auch Frührentner oder Teilnehmer von Rehabilitationsprogrammen sind aus dem Kreis der Arbeitslosen ausgeschlossen. Unter anderem auch deshalb sind die jeweils regierenden und an niedrigen Arbeitslosenzahlen interessierten Kreise so große Freunde von Langzeitstudenten und Umschulungsprogrammen aller Art.

Neben der reinen Zahl der Arbeitslosen gibt es auch noch die Arbeitslosen*quote*. Die braucht einen Nenner. Das ist die sogenannte „Erwerbsbevölkerung". Dazu zählen in Deutschland alle Bürger zwischen 15 und 65, die gegen Entgelt arbeiten oder arbeiten wollen und damit riskieren, arbeitslos zu werden. Beamte, obwohl eigentlich nicht dem Risiko der Arbeitslosigkeit ausgesetzt, zählen mit. Nicht dabei sind aber die Selbständigen, die Soldaten der Bundeswehr und alle Bauern (seit neuerem berechnet man auch Quoten auf Basis *aller* zivilen Erwerbspersonen, aber die sind wenig populär).

Auch diese Definition hat ihre Tücken. In den 80er Jahren des letzten Jahrhunderts hat man in England auf Anordnung von Margret Thatcher die zuvor im Nenner der Arbeitslosenquote nicht vertretenen englischen Staatsangestellten in den Nenner aufgenommen. Darauf fiel die englische Arbeitslosenquote über Nacht von 12 auf 10 Prozent. In einigen Medien wurde das als großer Erfolg gefeiert.

Arithmetisches Mittel

Das arithmetische Mittel ist die beliebteste Methode, Durchschnitte von was auch immer zu berechnen. Wenn ich gestern Abend vier Glas Wein getrunken habe, und vorgestern Abend zwei, macht das im Durchschnitt drei. Diese drei Glas Wein sind das arithmetische Mittel aus zwei und vier.

Dieser Durchschnitt ist nicht der einzig mögliche (siehe die Stichwortartikel Median sowie geometrisches und harmonisches Mittel), aber in der Regel der beste. Die Werte, deren Durchschnitt gesucht ist, werden aufsummiert, durch