Der Blick hinter das Buch: Zur Behandlung des kritischen Lesens im Unterricht zur Statistik[1]

Andrew Gelman, Columbia

[1] Original: „Going beyond the book: towards critical reading in statistics teaching“ in Teaching Statistics 35 (2013) 2, S. 82–86.
Übersetzung: Andreas Eichler, Freiburg

Zusammenfassung: *Wir diskutieren drei Beispiele unserer eigenen Lehrpraxis, die gute Lerngelegenheiten durch die kritische Betrachtung von Beispielen aus Büchern ermöglichen. Tatsächlich enthalten einflussreiche, weithin beachtete Bücher Beispiele mit Hilfe derer ohne zu großen Zusatzaufwand viel gelernt werden kann.*

1 Einleitung

Ein Mittel, den Stochastikunterricht aufzuwerten, kann in dem Ansatz bestehen, Beispiele aus Büchern aufzunehmen und zu diesen weitere Daten zu erheben, in den Büchern zitierte Artikel zu betrachten und schließlich ergänzende Datenanalysen auszuführen. Das alleine ist sicher keine große Neuigkeit. Was aber neu sein kann, ist die hier vorgestellte Einsicht, wie einfach die genannten Zusätze realisiert werden können: Gerade einflussreiche und hoch geschätzte Bücher umfassen Beispiele, an denen mit geringem Aufwand viel gelernt werden kann.

Wir betrachten hier drei Beispiele, die wir im eigenen Unterricht eingesetzt haben:

- Ein Lehrbuch, das uns veranlasst hat, sorgfältig über kategoriale und stetige Merkmale nachzudenken.
- Ein populärwissenschaftliches Buch, das Studien zum Zusammenhang von Menstruation und dem Risiko, einen Verkehrsunfall zu erleiden, fehlerhaft berichtet.
- Eine Monographie zu den Grundlagen der Wahrscheinlichkeitsrechnung, das die statistisch unauffällige Schwankung von Geschlechteranteilen überschätzt.

2 Kategorial oder stetig?

Das Lehrbuch *Mind on Statistics* von Jessica Utts und Robert Heackard (2001) ist exellent und enthält eine Fülle von Beispielen für alle Klassenstufen. Eine interessante Art und Weise, mit guten Lehrbüchern umzugehen, ist es, die dort behandelten Beispiele vertieft zu untersuchen. Beispielsweise werden in diesem Buch relativ zu Anfang kontinuierliche und kategoriale Merkmale behandelt. Dort werden die folgenden Merkmale als kategorial bezeichnet: Händigkeit (linkshändig, rechtshändig), der regelmäßige Kirchenbesuch (ja, nein), die Einstellung zur Legalisierung von Marihuana (ja, nein, unentschieden) und die Augenfarbe (braun, blau, grün oder nussbraun). Wechselt man aber die Perspektive, so können drei der vier Merkmale auch als kontinuierlich bzw. stetig verstanden werden.

Diese Behauptung ist am deutlichsten, wenn man die Händigkeit betrachtet, die Utts und Heckard mit den Ausprägungen links- und rechtshändig beschreiben. Dieses Merkmal kann aber ebenso als kontinuierlich verstanden werden, wie wir in Abb. 1 illustrieren, deren Daten auf einer bei Schülern erhobenen Stichprobe basieren (systematischer erhobene Daten haben ähnlich Resultate ergeben, vgl. Oldfield 1971). Wie im Histogramm zu sehen befinden sich viele Menschen hinsichtlich der Händigkeit zwischen den beiden Extremen. Das Histogramm in Abb. 2 zeigt dagegen, dass Schüler die Verteilung der Händigkeit bimodal und im Wesentlichen als diskret einschätzen. Diese nicht ungewöhnliche Fehlvorstellung bezogen auf die Händigkeit ist ein gutes Beispiel für ein kontinuierliches Merkmal, das häufig als diskret erachtet wird.

Ähnliche Betrachtungen sind auch bei anderen Merkmalen möglich, die Utts und Heckard nennen: Der Kirchbesuch kann etwa durch die Anzahl von Kirchbesuchen pro Jahr metrisch gemessen werden, was sogar informativer ist, als die Ausprägungen ja und nein bei einem kategorialen Merkmal. Etwa hat die American National Elections Study (http://www.electionstudies.org) gefragt, „Wie oft besuchen Sie religiöse Veranstaltungen, ausgenommen Hochzeiten und Beerdigungen?“, und enthielt fünf Arten von Antworten: Mehr als einmal pro Woche, einmal pro Woche, mehr als einmal pro Monat, einige Male im Jahr und nie. Schließlich kann auch die Meinung zur Legalisierung von Marihuana mit 0, 1 und 0,5 kodiert werden, während Zwischenwerte durch detailliertere Fragen identifiziert werden könnten wie etwa zu der Einstellung zur medizinischen Verwendung von Marihuana oder zu der Straffrage.

Der wesentliche Grund, diese Art von Diskussionen im Unterricht zu initiieren, ist nicht zu behaupten, dass etwa das Merkmal Kirchgang prinzipiell diskret oder stetig ist. Vielmehr geht es darum, Schülerinnen und Schüler für die Modellierung der Realität durch

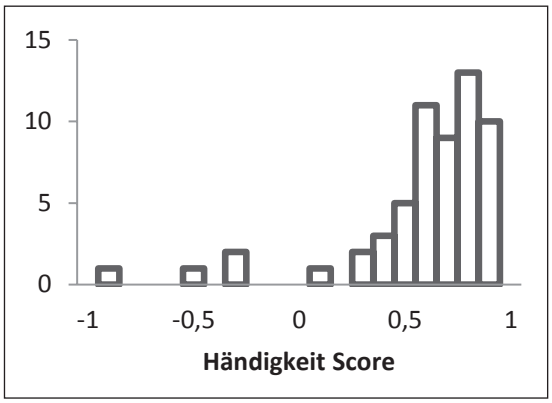

Abb. 1: Händigkeit kann mit einem 10-Item-Fragebogen gemessen werden, um einen im Prinzip stetigen Score zu ermitteln, der zwischen –1 (Linkshänder) und 1 (Rechtshänder) liegt. Schülerinnen und Schüler haben diesen Fragebogen ausgefüllt.

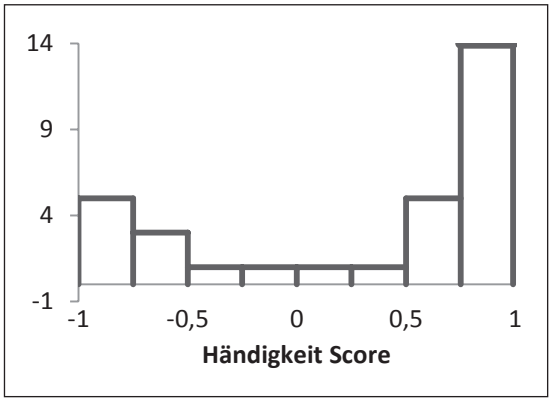

Abb. 2: Die Schülerinnen und Schüler wurden weiter gebeten, die von ihnen erwartete Verteilung der Händigkeit zu skizzieren. Die von diesen erwartete, oben zu sehende Bimodalität entspricht (Abb. 1) offenbar nicht der Realität.

Messung zu sensibilisieren. Wir halten es ebenfalls für sinnvoll, solcher Art Diskussionen durch eigenen Datensammlungen (Abb. 1) und Recherche, zum Beispiel der National Elections Studie anzureichern.

3 Der Graph, den es nicht gibt

Vor 15 Jahren, als ich ein Einführungskurs zur Statistik plante, erinnerte ich mich an die enthusiastische Rezension von Sills (1986) zu der 6. Auflage des Buchs von Hans Zeisel (1985) mit dem Titel *Say it with figures*. Ich kaufte das Buch und überflog es, um mögliche Beispiele für meinen Kurs zu finden. Dabei entdeckte ich zwei Skkizzen, die in Abb. 3 rekonstruiert sind. Beide Kurven repräsentieren Daten zu Krankenhauseinliefrungen von Frauen vor der Menopause, die in Verkehrsunfällen verwickelt wurden. Die jeweils linke der sich überlappenden Kurven zeigt die Unfälle, die sich vor der Menstruation ereignet haben, die rechte diejenigen nach der Menstruation.

Abb. 3: Skizze aus Zeisel (1985). Dieser schreibt: „Wenn die Häufigkeit von Unfällen bezogen auf die Zeit der Menstruation dargestellt wird, so entsteht ein überraschendes Bild (linker Graph). Es zeigen sich zwei getrennte Kurven (rechter Graph), einer für Frauen, die bereits Kinder haben, und einer für kinderlose Frauen. Die eine Gruppe hat das größte Unfallrisiko vor, die andere Gruppe nach der Menstruation.“ Dies ist zu vergleichen mit den Daten in Abb. 4.

Das Beispiel schien mir hoch geeignet für meinen Kurs. Da ich annahm, dass eine eigene Datensammlung das Beispiel noch verbessern würde, recherchierte ich in der Bücherei und fand die Arbeit von Katharina Dalton (1960). Die Daten aus dieser Arbeit sind in Abb. 4 dargestellt. Diese sind nicht einmal annähernd ähnlich zu den in Zeisel (1985) gegebenen Daten. Die skizzierten Wahrscheinlichkeitsdichten sollen die gesamte Wahrscheinlichkeitsmasse zu Unfällen direkt vor und nach der Menstruation repräsentieren, tatsächlich zeigen sie aber nur die Hälfte der Unfälle. Genauer ausgedrückt zeigen die skizzierte Dichten zwei Modalwerte mit einem Tal in der Mitte, während die Daten (Abb. 4) kein solches Tal implizieren. Ebenso ähneln auch die beiden glockenförmigen Kurven in Abb. 3 auf der rechten Seite nicht der korrespondierenden Darstellung in Abb. 4 unten.

Das Ergebnis von Dalton (1960) ist angemessen in einem Artikel des Magazins *Time* vom 28. November 1960 zusammengefasst: „In vier Krankenhäusern in London befragte Dr. Dalton 84 weibliche Unfallopfer mit einer Altersspanne von 15 bis 55. Alle Frauen hatten einen normalen 28tägigen Menstruationszyklus. Ihre Untersuchung ergab: 52 Prozent der Unfälle ereigneten sich mit einem Abstand von 4 Tagen zum Beginn der Menstruation. Unter der Hypothese, dass die Unfälle in dem Zyklus gleichverteilt wären, wären aber in der genannten Zeitspanne nur 28,5 Prozent der Unfälle zu erwarten gewesen. Kinderlose Frauen, so stellte Dr. Dalton fest, scheinen direkt vor der Menstruation ungewöhnlich deutlich unfallgefährdet zu sein, während Frauen mit Kindern während der gesamten Zeit kurz bevor und während

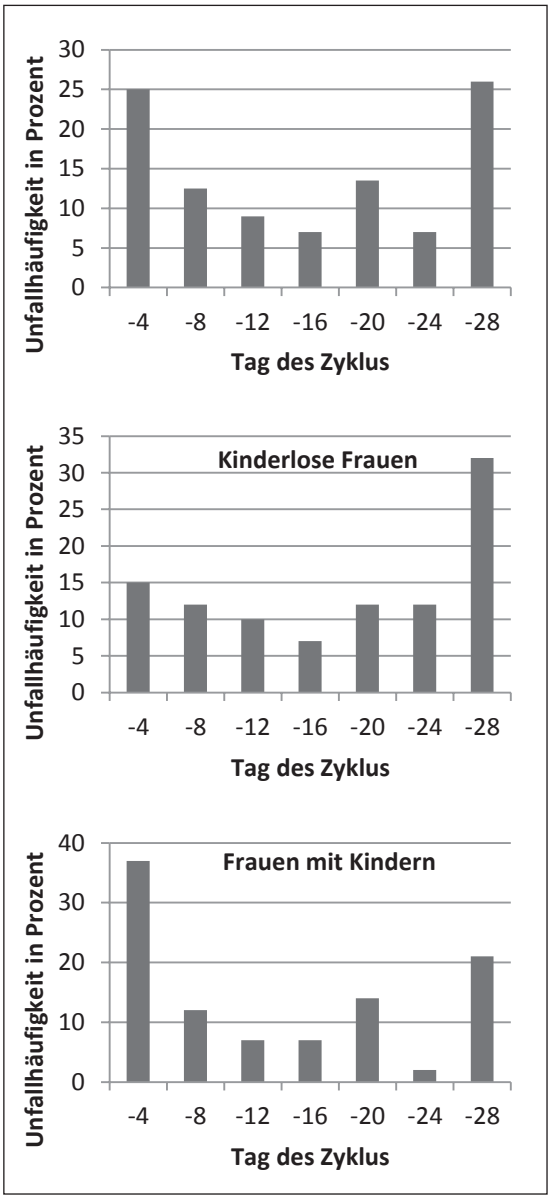

Abb. 4: Graphen nach Dalton (1960) mit den Rohdaten zu Unfällen im Menstruations-Zyklus. Die Graphen haben keine Ähnlichkeit mit der Skizze in Abb. 3, die aus dem Buch von Zeisel stammt. Viele Unfälle passen nicht zu den Modalwerten in Abb. 3. Zudem gibt es in diesen Daten kein Tal zwischen zwei Modalwerten.

der Menstruation gefährdet sind.“ (Übersetzung des Herausgebers)

Was hier relevant für die Betrachtungen ist, dass die tatsächliche Ergebnisse in diesem Kontext in dem Buch von Zeisel nicht präzise wiedergegeben sind. Eine aus heutiger Sicht skurrile Nebennote ist zudem, dass in dem Artikel Dalton mit der Überlegung zitiert ist, man könne darüber nachdenken, ob die Ergebnisse nicht die Verschreibung von Beruhigungsmitteln vor der Menstruation als angebracht erscheinen ließen.

Ich vermute, dass Zeisel die Ergebnisse oder den zitierten Artikel kannte und ihr Potenzial als Unterrichtsbeispiel sah. Möglicherweise hat er dann zu voreilig die Daten in den Skizzen (Abb. 3) zusammengefasst und die meisten Unfälle vernachlässigt, die sich nicht unmittelbar vor oder nach der Menstruation ereigneten und dadurch irrtümlicherweise ein Tal zwischen zwei Modalwerten in der Skizze erzeugte. Möglicherweise suchte er auch direkt nach einem solchen bimodalen Modell und beachtete die tatsächlichen Daten nur oberflächlich. Wie auch immer, für unsere Schülerinnen und Schüler ist das ein Glücksfall, durch den sie erfahren können, wie leicht man die Ergebnisse einer Studie fehlinterpretieren kann. Betont werden soll aber hier noch einmal: Gerade weil Zeisels Buch so hervorragend geschrieben ist, nehmen wir uns heraus, dieses Beispiel hier herauszustellen.

Für Lehrkräfte kann es eminent wichtig sein, sekundärer Daten auf den Grund zu gehen und nach Verbesserungen zu suchen. Für Schülerinnen und Schüler geht es darum, selbst bei vertrauenswürdigen Quellen den skeptischen Blick auf sekundäre Dateninterpretationen nicht zu verlieren.

4 Muster im Rauschen

Das Buch *Probability, Statistics and Truth* von Richard von Mises (1957; in der deutschen Ausgabe 1928; Anm. des Hrsg.) ist eine wichtige Arbeit über die Fundamente der Wahrscheinlichkeitsrechnung und enthält den Versuch Axiome aus unendlichen Serien von Zufallsversuchen abzuleiten. Diese Arbeit war lange sehr einflussreich und das sowohl in der Stochastik (vgl. etwa Wald 1939 oder Good 1958 für Reaktionen aus einer klassischen und einer bayesianischen Perspektive) als auch in der Philosophie (z. B. Gillies 2000, der die Ideen von Mises mit denen von Karl Popper verbindet).

Ich kaufte mir dieses Buch vor einigen Jahren und wurde besonders inspiriert von dem Kapitel zu Anwendungen der Stochastik. Dort werden die Anteile der Geschlechter zur Illustration der Binomialverteilung verwendet. Von Mises berichtet von Jungenanteilen in Wien innerhalb der 24 Monaten der Jahre 1907 und 1908 und von der in diesem Zeitraum unerwartet geringen Streuung. So sei der Mittelwert der Jungenanteile 0,51433, die mit n normierte Varianz 0,0000533. Die erwartete Varianz berechnete er durch

$$\frac{23}{24} \cdot \frac{p(1-p)}{n} = 0{,}0000613\,,$$

wobei hier n die Geburtenanzahl von ungefähr 3900 pro Monat und p mit 0,514 geschätzt ist. Von Mises schreibt anschließend: „Die tatsächliche Streuung ist kleiner als die theoretisch zu erwartende. In anderen Untersuchungen zu der Jungenrate ist der Wert des Lexis-Quotienten näher an 1. Daher könnte man nach einer Erklärung für die auffällige Streuung in diesem speziellen Fall suchen.“ (Übersetzung des Hrsg.) Von Mises fährt dann fort, indem er die Streuung der Geschlechteranteile mit unterschiedlichen Rassengruppen sowie sozioökonomischen Gruppen erklärt.

Allerdings muss die Tatsache, dass die Varianz unter der Annahme eines konstanten Mittelwerts des Geschlechteranteils geringer ist als erwartet, nicht notwendigerweise statistisch auffällig sein. Das kann man etwa mit dem chi-Quadrat-Test untersuchen – das ist eine in den Sekundarstufe unübliche Methode, die aber ähnlich zu anderen Testmethoden ist. Die Nullhypothese in diesem Test wird aus der Annahme abgeleitet, dass die Geburtenanzahl pro Monat binomialverteilt ist mit konstantem p. Weiter sei die Teststatistik durch $\frac{(n-1)s^2}{\sigma^2}$ konstruiert, also durch den Vergleich der Varianz in der Bevölkerung und einer speziellen Stichprobe.

Für die 24 Monate wird die empirische Varianz durch

$$s^2 = \frac{24}{23}\,(0{,}0000533)$$

geschätzt auf der Basis von 23 Freiheitsgraden. Diese Varianz wird verglichen mit der theoretischen Varianz, für die sich unter der Annahme eines konstanten Geschlechteranteils ergeben hatte:

$$\sigma^2 = \frac{24}{23} \cdot 0{,}0000613$$

Da die Testgröße $\frac{23}{24} \cdot \frac{s^2}{\sigma^2}$ einer chi-Quadrat-Verteilung mit 23 Freiheitsgraden genügt, hat ein Quotient von mindestens 0,869 einen p-Wert von 0,36. D. h. in mehr als einem Drittel der Fälle würde man einen

Wert des Quotienten von 0,869 oder kleiner rein zufallsbedingt beobachten.

Damit ist es aber auch unnötig, nach einem Grund für die Diskrepanz der Varianzen zu suchen. Das gilt insbesondere dann, wenn tatsächlich die Jungen- und Mädchenanteile mit einer Binomialverteilung modellierbar sind. Man könnte noch ergänzen, dass von Mises bezogen auf die Schulpraxis mit dem Vergleich der Varianzen einen ungünstigen Weg einschlägt. Dort wären eher die besser in der Originalskala interpretierbaren Standardabweichungen sinnvoll.

Von Mises ist sicherlich nicht allein beim Überinterpretieren von Geburtsdaten. So gibt es eine lange Tradition, Geburtsdaten zu analysieren, obwohl es bisher keine überzeugenden Hinweise etwa zu Jungen- oder Mädchen-Runs in Familien gibt. Ebenso wenig scheinen die Anteile von Mädchen oder Jungen – außer bei sehr ungewöhnlichen Rahmenbedingungen – im obigen Sinne auffällig zu sein (vgl. Freese & Powell 2001, Das Gupta 2005, Gelman 2007 für weitere Argumente gegen die Überinterpretation von statistischen Schwankungen in Geschlechteranteilen). Damit ergibt dieses Beispiel nicht nur die Möglichkeit, die wichtige Einsicht in die statistische Überprüfung von Varianzen, sondern eröffnet ebenso den kritischen Blick auf Fehlinterpretationen von Statistiken. Anmerkung: Von Mises nutzte für seine Berechnungen inkorrekterweise die Formel für die Populationsvarianz. Wir laden alle Leser ein, der Originalquelle zu folgen, die eine Varianz von 0,0000394 angibt, was einen ordentlichen *p*-Wert von 0,10 im Gegensatz zu dem oben angegebenen Wert von 0,36 ergäbe.

Dass so etwa in einem hochgeachteten Buch passiert zeigt bloß, dass auch ein Standard-Test wie der chi-Quadrat Test für die Über-Streuung nicht als selbstverständlich betrachtet werden darf. In gleicher Weise gibt es ja auch die Erkenntnis, dass der große Francis Galton fehlerhafte Berechnungen im Rahmen der Normalverteilung durchgeführt hat (Gelman 2006, Wainer 2007).

Diskussion

Für sich gesehen, haben alle angeführten Beispiele wenig Wert. Tatsächlich wird ja niemand ein Lehrbuch zur Statistik dafür verwenden, um etwas über Händigkeit, Menstruation oder Geschlechteranteile zu lernen. Dennoch ist es interessant, dass in diesen Büchern Beispiele mit interessanten Datenmustern einer nachträglichen Überprüfung nicht Stand halten können. In dem einen Beispiel hat etwa das Zurückgehen auf die Originaldaten ergeben, dass Zeisel bei der Übernahme von Daten in eine Grafik die Aussagen der Originalstudie fehlinterpretiert hat. Eine statistische Re-Analyse des Beispiels aus dem Buch von von Mises hat ergeben, dass die schon für sich fragwürdige Zuschreibung von ethnischen Gründen für das Auftreten von Geschlechteranteilen schon aus dem Grund nicht haltbar sind, da die empirisch aufgetretenen Unterschiede statistisch nicht signifikant sind. Schließlich hat uns das Beispiel von Utts und Heckard zur Händigkeit und dem Kirchbesuch zu einer eigenen Datensammlung und im Ergebnis zu neuen Erklärungen geführt.

Gute Lehre ist dann gegeben, wenn Schülerinnen und Schüler Berichte und wissenschaftliche Artikel, die einen statistischen Inhalt haben, kritisch hinterfragen können (Gelman & Nolan 2002). In diesem Beitrag ist die Empfehlung, diesen Blick auch bei den Büchern, die wir in der Lehre verwenden, beizubehalten. Vieles kann dadurch gelernt werden, wenn wir Re-Analysen von Daten durchführen und Originaldaten wie deren sekundären Analysen sehr sorgfältig untersuchen. All das kann unsere Lehre bereichern, selbst wenn wir dabei die von uns favorisierten Bücher in Zweifel ziehen.

Danksagung

Wir danken Ji Meng Loh, Martin Lindquist, Roger Johnson für die hilfreichen Kommentare und der U.S. National Science Foundation, den National Instititutes of Health und dem Columbia University Applied Statistics Center für die finanzielle Unterstützung.

Literatur

Dalton, K. (1960): Menstruation and accidents. In: *British Medical Journal, 2,* S. 1425–1426.

Das Gupta, M. (2005): Explaining Asia's „missing woman": A new look at the data. In: *Population and Development Review, 31(3),* S. 529–535.

Freese, J. & Powell, B. (2001): Making love out of nothing at all? Null findings and the Trivers-Willard hypothesis. In: *American Journal of Socialogy, 106(6),* S. 1776–1788.

Gelman, A. (2006): Galton was a hero to most. Statistical modeling, causal inference, and social science blog, 23. October 2009, http:///www.stat.columbia.Edu/~gelman/blog.

Gelman, A. (2007): Letter to the editor regarding some papers of Dr. Satoshi Kanazawa. In: *Journal of Theoretical Biology, 245(3),* S. 597–599.

Gelman, A. & Nolan, D. (2002): *Teaching statistics: A bag of tricks.* Oxford: Oxford University Press.

Gillies, D. (2000): *Philosophical theories of probability.* London: Routledge.

Good, I. J. (1958): Review of *Probability, statistics and truth* von R. von Mises. In: *Journal of the Roiyal Statistical Society, Series A, 121(2),* S. 238–240.

Oldfield, R. C. (1971): The assessment and analysis of handedness: The Edinburgh inventory. In: *Neuropsychologia, 9(1),* S. 97–113.

Sills, D. L. (1986): Review of *Say it with figures* by Hans Zeisel. In: *Journal of the American Statistical Association, 81(393),* S. 257.

Utts, J. M. & Heckard, R. F. (2001): *Mind on Statistics.* Pacific Grove, CA: Duxbury.

Von Mises, R. (1957): *Probability, statistics, and truth.* New York: Dover.

Wainer, H. (2007): Galton's normal is too platykurtic. In: *Chance, 20(2),* S. 57–58.

Wald, A. (1939). Review of *Probability, statistics, and truth* von R. von Mises. In: *Journal of the American Statisrical Association, 34(2007),* S. 591–592.

Zeisel, H. (1985). *Say it with figures.* New York: Harper and Row.

Anschrift des Verfassers

Andrew Gelman
Columbia University
gelman@stat.columbia.edu