

Statistische Grundlagen

im Zusammenhang mit der [Veröffentlichung](#):

Schwarz C, Kratochvil E, Pilger A, Kuster N, Adlkofer F, Rüdiger HW:
Radiofrequency electromagnetic fields (UMTS, 1,950 MHz) induce genotoxic effects in vitro in human fibroblasts but not in lymphocytes.

und den [Kommentaren](#) hierzu:

Lerchl, A:

Comments on "Radiofrequency electromagnetic fields (UMTS, 1,950 MHz) induce genotoxic effects in vitro in human fibroblasts but not in lymphocytes" by Schwarz et al. (Int Arch Occup Environ Health 2008: doi: 10.1007/s00420-008-0305-5).

sowie der Berichterstattung im [Laborjournal](#):

Bär, S:

Wer wird hier von der Industrie bezahlt?

(Hinweis: durch Klicken auf die Links kommen Sie zu den Artikeln)

Version 1

© Alexander Lerchl 2008.

Dieser Text darf nicht, auch nicht auszugsweise, zu kommerziellen Zwecken in irgendeiner Weise wiedergegeben oder vervielfältigt werden. Zu nicht-kommerziellen Zwecken darf er ohne Beschränkung verteilt werden, wenn die Quelle und der Autor korrekt zitiert sind.

Quelle: www.mobilfunkdebatte.de

Für Anregungen und Kritik bin ich stets dankbar: a.lerchl@jacobs-university.de

Bremen, 2. Mai 2008

1. Einleitung

Bei der Diskussion um die auf der ersten Seite genannten Studien kommen statistischen und methodischen Argumenten große Bedeutung zu, die für Laien schwer verständlich sind. Diese kurze Darstellung soll helfen, auch Laien zu ermöglichen, die Diskussion nachzuvollziehen und sich ein eigenes, fundiertes Bild zu machen.

Wenn Sie bereits wissen, was Standardfehler, Standardabweichung, Variationskoeffizienten, parametrische und nicht-parametrische Tests usw. sind, können Sie die Kapitel 2 und 3 überspringen. Wenn das nicht der Fall ist, sollten Sie diese Kapitel besonders gründlich lesen.

Hinweis an Experten: Ich habe den Text bewusst so verfasst, dass er für Laien verständlich ist. Daher sehen Sie mir nach, wenn ich auf einige wichtige, aber sehr abstrakte und hier auch unwesentliche Sachverhalte nicht verweise.

2. Begriffsbestimmungen

Statistische Verfahren lassen sich grob in zwei Kategorien unterteilen: die beschreibende Statistik (**deskriptive S.**) und die **analytische** Statistik. Beispiele für die deskriptive Statistik sind die Zahlen, die regelmäßig von statistischen Landes- oder Bundesbehörden kommen, analytische Statistiken dienen zum Testen auf Unterschiede, Verteilungsmuster und dergleichen. In den Naturwissenschaften sind statistische Methoden wichtiger Bestandteil von Untersuchungen.

Beispiel 1: wir wollen wissen, wie schwer und wie groß Kinder im Alter von 10 bis 14 Jahren sind. Ziel ist es herauszufinden, ob die Kinder im Vergleich zu früheren Jahren und im Vergleich zu anderen Ländern größer oder kleiner, schwerer oder leichter sind. Dazu können wir entweder **alle** Kinder hierzulande vermessen (sog. **Grundgesamtheit**) oder eine **Stichprobe** wählen, um die **wahren Werte** (nämlich der Grundgesamtheit) **abzuschätzen**. Völlig klar ist, dass wir mit wenigen Kindern keine ausreichenden Daten bekommen, die Resultate sind dann **nicht belastbar**. Erst mit einer **repräsentativen** und ausreichend großen Stichprobe, die die Grundgesamtheit möglichst gut repräsentiert, können wir Daten gewinnen, die mit **hinreichender Wahrscheinlichkeit** Rückschlüsse auf die Werte aller Kinder zulassen. Bei solchen Erhebungen sind Stichprobengrößen im vierstelligen Bereich üblich.

Beispiel 2: in einer **Zellkultur**, in der sich zigtausende Zellen befinden, soll die Aktivität eines bestimmten Enzyms getestet und mit einer anderen Zellkultur verglichen werden. Ein Testverfahren ist verfügbar, das die Messung der Enzym-Aktivität in jeder einzelnen Zelle zulässt. Wiederum werden Stichproben aus beiden Zellkulturen entnommen, z.B. 500 Zellen. Dies wird als ausreichend angesehen, um Aussagen über alle Zellen der je-

weiligen Kultur zu erhalten. Anschließend sollen die Ergebnisse der Zellkulturen in einem Testverfahren dahingehend geprüft werden, ob sie sich **statistisch** unterscheiden.

Wichtig an diesen Beispielen ist, dass wir mit den Ergebnissen der Stichprobe nur Aussagen treffen können, die mit einer gewissen **Wahrscheinlichkeit** für alle Kinder / Zellen gelten. Je größer die Stichprobe, umso höher die Wahrscheinlichkeit, dass die erhobenen Daten mit den realen Daten übereinstimmen. Deswegen spricht man in der Statistik häufig von **Schätzungen**. An dieser Stelle sind jetzt einige Begriffe zu erläutern, die wir für die weitere Diskussion benötigen:

- Mittelwert: Summe aller Messwerte, geteilt durch die Anzahl Messungen (N).
- Standardabweichung: Maß für die Streuung der Einzeldaten um den Mittelwert. Mathematisch ist dies die Quadratwurzel der Summe der Abweichungen der Einzelwerte vom Mittelwert, geteilt durch die Anzahl Messungen -1 ($N-1$).
- Standardfehler: Maß für die Streuung der Einzeldaten um den Mittelwert, allerdings berechnet aus dem Quotienten der Standardabweichung und der Quadratwurzel der Anzahl Messungen.
- Variationskoeffizient: Relatives Maß (in %) für die Streuung der Einzelwerte um den Mittelwert, berechnet aus dem Quotienten aus Standardabweichung und Mittelwert.

Wichtig ist, dass die Standardabweichung nur unwesentlich von der Anzahl Messungen abhängt (wegen $(N-1)$ im Nenner), der Standardfehler aber umso kleiner wird, je mehr Messungen vorliegen. Da diese Begriffe für Laien schwer verständlich sind, sollen sie an weiteren Beispielen verdeutlicht werden.

Angenommen, wir haben folgende 10 Daten gemessen: 8, 10, 12, 12, 10, 8, 10, 12, 8, 10. Der Mittelwert ist genau 10, die Standardabweichung ist 1,63 und der Standardfehler 0,52. Der Variationskoeffizient beträgt 16,3%.

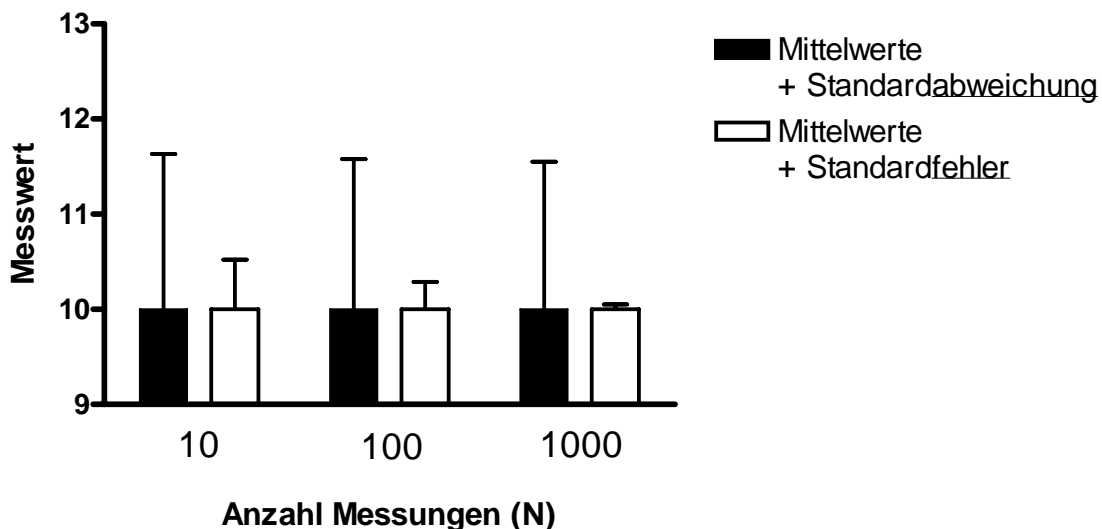
Nun wiederholen wir die Messungen zwei weitere Male und lassen die nun 30 Zahlen in die Analyse eingehen: 8, 10, 12, 12, 10, 8, 10, 12, 8, 10, 8, 10, 12, 12, 10, 8, 10, 12, 8, 10, 8, 10, 12, 12, 10, 8, 10, 12, 8, 10. Der Mittelwert ist zufällig wieder genau 10, die Standardabweichung ist mit 1,58 etwas geringer, aber der Standardfehler ist nur noch 0,29. Der Variationskoeffizient beträgt 15,8%.

Das Ganze wiederholen wir noch öfter, und erhalten bei $N=100$ die Standardabweichung von 1,56 und den Standardfehler von 0,16, bei $N=1000$ Messungen erhalten wir für die Standardabweichung den Wert von 1,55, während der Standardfehler nur noch bei 0,05 liegt.

Mit anderen Worten: mit zunehmender Anzahl Messungen verändert sich die Standardabweichung nur geringfügig, während der Standardfehler immer kleiner wird. Beide Angaben haben ihre Vor- und Nachteile, und oft ist es nur eine Geschmacks- oder Ansichtssache, ob nun der eine oder der andere Wert angegeben wird. In jedem Fall gehört aber zur Angabe eines Mittelwertes mit Standardfehler immer die Angabe der Anzahl Messwerte, da man als Leser ohne die Originaldaten nicht ermitteln kann, wie die Streuung der Daten wirklich war. In der Publikation von Schwarz et al. wird immer die Standardabweichung angegeben.

Zwischenfazit: Die Standardabweichung ist immer größer als der Standardfehler. Je mehr Beobachtungen, umso geringer der Standardfehler. In einem Bild dargestellt ergeben sich für die Daten bei gleicher Streuung unterschiedliche Eindrücke:

Gleiche Daten, unterschiedliche Darstellung



In der Publikation von Schwarz et al. wird an einer Stelle deutlich, dass die Autoren diese grundlegenden Sachverhalte nicht kennen, da sie behaupten, die Standardabweichungen würden durch die Analyse von 500 Zellen sehr gering werden („Due to the scoring of 500 cells, being about ten times the cells usually processed by computer-aided image analysis, standard deviations become very low.“). Dass sie aber in der Tat Standardabweichungen und nicht Standardfehler berechnet und in Tabelle 2 und in den Abbildungen dargestellt haben, ist unstrittig. Die Abbildungen und die Tabelle weisen auf außerordentlich geringe Standardabweichungen und damit Variationskoeffizienten hin, die immer unter 5% liegen (aus urheberrechtlichen Gründen kann ich die hier weder die Abbildungen noch die Tabellen komplett zeigen). Diese geringen Abweichungen sind für derart komplexe Experimente viel zu gering und widersprechen auch früheren Ergebnissen derselben Arbeitsgruppe. Ein sehr interessantes Einzelergebnis meiner Analysen ergab, dass die Variationen zwischen den Ergebnissen einzelner Experimente geringer waren als die Variationen in einem Einzelexperiment. Dies ist aus methodischen und logischen Gründen nicht möglich.

3. Statistische Tests

Um zwei oder mehr Mittelwerte aus Messungen zu vergleichen, gibt es eine ganze Reihe von statistischen Verfahren. Diese mit ihren Bedingungen hier aufzulisten ist nicht sinnvoll und würde mehr verwirren als nützen. Die beiden wichtigsten Grundtypen für Vergleiche im medizinischen und biologischen Bereich sind:

- Parametrische Tests: Hier werden Werte (Mittelwerte) mit ihren Standardabweichungen bzw. Standardfehlern und der Anzahl Beobachtungen als tatsächliche Werte miteinander verglichen. Je nachdem, ob und ggf. wie weit sich diese Werte überschneiden, ergeben die Tests eine **Wahrscheinlichkeit** (s.u.), ob sie statistisch signifikant unterschiedlich sind.
- Nicht-parametrische Tests: Hier werden nicht die Werte an sich miteinander verglichen, sondern sie werden der Größe nach sortiert (Rangfolge), um anschließend zu testen, ob z.B. die höchsten Werte in einer Gruppe sind und die niedrigsten in der anderen.

Bei beiden Verfahren wird am Ende ein Wert geliefert, der die **Irrtumswahrscheinlichkeit** dafür ist, dass der gefundene Unterschied tatsächlich **nicht** besteht (sog. „falscher Alarm“ oder statistischer Irrtum erster Art). Dieser Wert wird mit dem Buchstaben p (für engl. probability) bezeichnet. $p < 0,05$ bedeutet, dass diese Wahrscheinlichkeit 5% oder weniger beträgt, $p < 0,01$ steht entsprechend für 1% oder weniger usw. Je kleiner der Wert für p , umso „signifikanter“ ist der Unterschied. In der Wissenschaft gilt ein p -Wert von 0,05 oder kleiner als Schranke für eine statistische Signifikanz.

Beide Verfahren haben Vor- und Nachteile. Parametrische Tests sind häufig empfindlicher, was sog. Ausreißer angeht. Bei ihnen muss vor dem Testen genau geschaut werden, ob die zu vergleichenden Daten a) in etwa gleich streuen (sog. **Gleichheit der Varianzen**) und b) ob sie normalverteilt sind. Erst wenn beide Voraussetzungen erfüllt sind, können solche parametrischen Tests ohne weiteres durchgeführt werden.

Nicht-parametrische Tests haben den Vorteil, mit ihnen auch stark oder unterschiedlich streuende Messwerte miteinander vergleichen zu können, bei denen die Voraussetzungen für die parametrischen Tests also nicht gegeben sind. Sie gelten als „robust“, aber „konservativ“, weil sie eine bestimmte **Mindestanzahl von Messwerten** benötigen, um überhaupt signifikante Unterschiede aufzeigen zu können. Dieser Punkt ist von zentraler Bedeutung für die weitere Diskussion und soll daher genau betrachtet werden:

Nehmen wir an, wir vergleichen die Körpergrößen von 3 Mädchen und 3 Jungen gleichen Alters. Die Werte seien für die Mädchen 173 cm, 178 cm und 175 cm. Für die Jungen werden ermittelt: 179 cm, 183 cm und 185 cm. Alle Jungen sind also größer als die Mädchen, es gibt keine Überschneidung der Messwerte. Der parametrische Test (sog. t-Test)

liefert einen p-Wert von 0,0375, der Unterschied ist also auf dem 5%-Niveau **signifikant**. Der nicht-parametrische Test (sog. Mann-Whitney-Test) hingegen zeigt **keinen signifikanten Unterschied** an ($p=0,1$)! Das liegt daran, dass die Reihenfolge Junge-Junge-Junge-Mädchen-Mädchen-Mädchen auch Zufall sein kann. Erst bei 4 Jungen und 4 Mädchen in jeder Gruppe kann dieser Test ein signifikantes Ergebnis liefern, wenn alle Jungen größer sind als alle Mädchen. Wenn hingegen die Anzahl der Jungen (3) gleich bleibt, kann nur mit **mindestens 5** Mädchen, die alle kleiner sind, der nicht-parametrische Test ein signifikantes Resultat liefern. Bei 3 Jungen und 6 Mädchen liegt der kleinste p-Wert beim nicht-parametrischen Test bei 0,0238 (2.38%). Bei parametrischen Tests hingegen ist der kleinste p-Wert nicht definiert, er kann also z.B. bei $p < 0,0001$ (0,01% Irrtumswahrscheinlichkeit) oder noch niedriger liegen.

4. Die Statistik in Schwarz et al.

In der Studie wurden jeweils 4 Gruppen von Zellen untersucht, die entweder **schein-exponiert** waren (sich in der Expositionseinrichtung im Inkubator befanden, aber nicht exponiert wurden), **exponiert** wurden, als **Inkubator-Kontrolle** (im Inkubator, nicht aber in der Expositionsanlage) oder als **positive Kontrolle** dienten, indem sie starker UV-Strahlung ausgesetzt wurden. In die Auswertung (statistische Tests) gingen jeweils 3 bzw. 6 Werte ein, und zwar 3 für die (Mobilfunk-) exponierten Zellen und 6 für die kombinierten schein-exponierten Zellen und die Inkubator-Kontrollen. Der statistische Test war in jedem Fall **nicht-parametrisch**.

Es waren also, wie wir jetzt wissen, **p-Werte von allenfalls 0,0238** zu erwarten (im Text der Veröffentlichung ist immer „0.02“ zu lesen). Interessant ist dabei:

- Der eigentliche, interessante Unterschied wäre der zwischen exponierten und nicht-exponierten Zellen gewesen. Nur wäre bei diesem statistischen Vergleich bei 3 Daten pro Gruppe **niemals ein signifikantes Ergebnis** herausgekommen.
- Daher **mussten** die Werte der schein-exponierten Zellen und der Inkubator-Kontrollen **zusammengefasst** werden, um überhaupt zu signifikanten Ergebnissen zu kommen. Das war aber **nur dann möglich**, wenn sich diese Werte statistisch nicht unterschieden (ansonsten wäre keine Kombination möglich gewesen). Dies war überraschenderweise in jedem Fall gegeben (warum das überraschend ist, wurde im Laborjournal-Bericht und in meiner Publikation eingehend dargelegt).
- Der unempfindliche nicht-parametrische Test und die Kombination der Werte waren aber **gar nicht notwendig**, weil die Voraussetzungen für einen viel empfindlicheren parametrischen Test in jedem Fall gegeben waren (Gleichheit der Varianzen, Normalverteilung). Dann wären die Signifikanzen sogar extrem deutlich ($p < 0,001$) gewesen. Warum diese Analyse nicht gewählt wurde (wie seinerzeit bei der Studie zu Wirkungen von GSM 1800; Diem et al., 2005), ist unklar.