9. Logarithmische Normalverteilung

9.1 Transformationen

In Kapitel 8 haben wir gesehen, dass vor einem Signifikanztest geklärt werden muss, ob die Daten normalverteilt sind. In Medizin und Biologie werden auch Daten erhoben, die nicht der Vorstellung von einer Normalverteilung entsprechen. Neben anderen Formen treten gelegentlich linksgipfelige Verteilungen auf. Um solche Daten Signifikanztesten unterziehen zu können, die Normalverteilung voraussetzen, können wir versuchen, sie so zu transformieren, dass die transformierten Daten normalverteilt sind. Zu einer Transformation müssen wir die Daten in „irgendeiner" Form rechnerisch bearbeiten, etwa indem wir sie mit einem Exponenten potenzieren oder radizieren. Welcher Algorithmus zum Erfolg, also dazu führt, dass die so bearbeiteten Daten dann normalverteilt sind, das hängt von der Datenstruktur ab und muss in der Regel empirisch geprüft werden.

Wir wollen am Beispiel 1 eine solche Transformation durchführen.

Beispiel 1 Nach Literaturangaben (Ramm, Hoffmann; Biomathematik, Enke-Verlag) ist die Anzahl der Leukozyten im menschlichen peripheren Blut linksgipfelig verteilt. In Ermangelung eines ausreichend großen Datenbestandes benutze wir hier einen fiktiven Datensatz von Leukozytenzahlen. Die Daten liegen klassiert mit der Klassenbreite 0,5 vor. Es soll geprüft werden, ob sie linksgipfelig verteilt sind. Wenn ja, dann sollen sie so transformiert werden, dass sie danach normalverteilt sind.

KM	4.5	5.0	5.5	6	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5
H	1	2	4	6	9	11	8	6	5	4	3	2	2	1	1

Tabelle 1, KM = Klassenmitte, alle Werte * 10^9 = Zellen/L; H = absolute Häufigkeit

Wenn wir die absolute Häufigkeit der Leukozytenzahlen gegen die Klassenmitten auftragen, dann resultiert eine linksgipfelige Verteilung (Abb.1). Das gleiche Ergebnis würden wir mit den relativen Häufigkeiten erhalten. Die Verteilung ist asymmetrisch, es liegt also keine Normalverteilung vor.

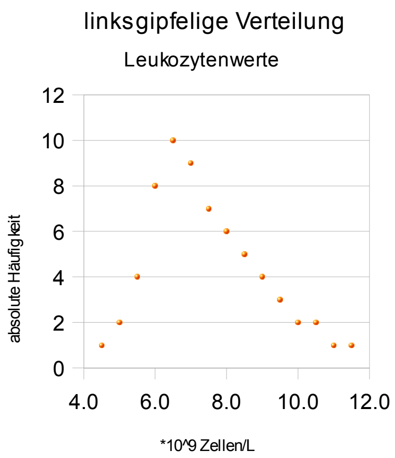

Abb. 1

Im Hinblick auf die mögliche Anwendung eines Signifikanztests wollen wir versuchen, die vorliegenden Daten so zu transformieren, das die transformierten Daten nomalverteilt sind.

9.2. Logarithmische Transformation links-gipfelig verteilter Daten

Bei linksgipfelig verteilten Daten hat es sich als erfolgreich erwiesen, die Daten auf einer logarithmisch geteilten Abszisse oder die Logarithmen der Daten auf einer metrischen Abszisse aufzutragen.

9.2.1 Die logarithmische Skala

Wir wollen die Daten zunächst in einem Koordinatensystem darstellen, dessen Abszisse logarithmisch geteilt ist. Dazu haben wir drei Möglichkeiten.

1. Wir nutzen die Erstellung einer logarithmischen Abszisse in einem Tabellenkalkulationssystem.

2. Wir verwenden das im Handel (Schleicher und Schüll) erhältliche halblogarithmische Netzpapier, dessen Abszisse über mehrere Zehnerpotenzen logarithmisch geteilt ist. Abb.2.

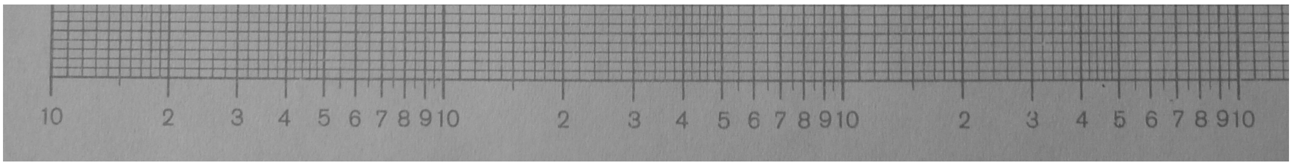

Abb.2

3. Wir erstellen selber eine logarithmisch geteilte Abszisse.

Zu Punkt 1. Mit dem Programm „OpenOfficeCalc“ erhalten wir bei Gegenüberstellung der absoluten Häufigkeiten gegen die Zählwerte auf logarithmischer Abszisse die Darstellung in Abb.3.

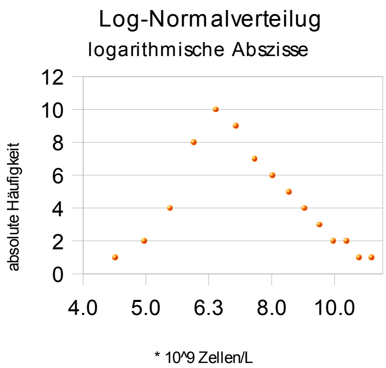

Abb.3

Der in Abb.1 links liegende Gipfel ist deutlich zur Mitte verschoben. Die Graphik erscheint hinreichend symmetrisch um die Daten als logarithmisch normalverteilt ansehen zu können. Wir sprechen dann von einer log-Normalverteilung und sagen dann „Die Leukozytenzahlen sind log-normalverteilt.“

Zu Punkt 2. Die Darstellung auf dem handelsüblichen Netzpapier stellen wir hier nicht dar.

Zu Punkt 3. Für den Fall, dass die Möglichkeiten 1 und 2 ausfallen, zeigen wir hier, wie wir eine logarithmische Abszisse selber erstellen können.

9.2.2 Konstruktion einer logarithmischen Skala

Ganz einfach wäre es, die Skala von einem Rechenschieber, dessen Skalen logarithmisch geteilt sind, zu übertragen. Aber wer hat schon einen Rechenschieber? Also machen wir es mit dem Taschenrechner, dem PC oder einer Logarithmentafel. Wir verwenden Logarithmen zur Basis 10, die dekadischen oder Briggschen Logarithmen. Auf dem Taschenrechner können wir sie mit der Taste ***log*** aufrufen. (Die Taste ***ln*** generiert die natürlichen Logarithmen, deren Basis die Eulersche Zahl e = 2,7182... ist.)

Wir gehen wie folgt vor:

1. Wir erstellen (am besten auf Millimeterpapier) eine metrisch äquidistant geteilte Hilfsskala mit den Zahlen 0 bis 1 (Abb.4)

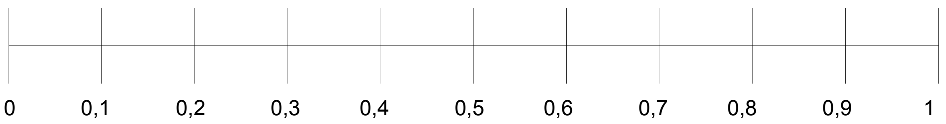

Abb.4

2. Dann generieren wir die Logarithmen für die Zahlen 1 bis 10 mit dem Taschenrechner. Sie stehen rot in der Tabelle 2.

Zahl	1	2	3	4	5	6	7	8	9	10
log Zahl	0	0,301	0,477	0,602	0,699	0,778	0,845	0,903	0,954	1

Tabelle 2

3. Wir tragen die Zahlen 1 bis 10 über der äquidistanten Skala nach folgenden Angaben ein: Wir tragen die 1 über deren Logarithmus (= 0) an der Skala ein. Die 2 tragen wir über ihren Logarithmus 0,310 ein, und die 3 über 0,477 usw.
Wenn die metrische Skalierung von 0 bis 1 in Abb.5 mit dem Inkrement 0,01 unterteilt wäre, dann könnten wir die Zahlen sicherer Positionieren als in Abb.5. Da diese Unterteilung fehlt, können wir den Ort 0,477 für die 3 nur schätzen. Daher die Empfehlung für das Millimeterpapier.

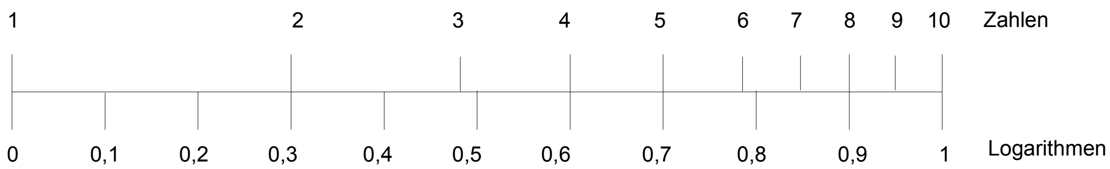

Abb.5

4. Wenn wir nun die untere Hilfsskala entfernen, dann bleibt eine Skala, auf der die Zahlen 1 bis 10 in logarithmischen Abständen aufgetragen sind. So entsteht Abb 6. Die Skala ist asymmetrisch und nicht äquidistant. Beachten Sie, dass die Abstände zwischen zwei Zahlen auch wieder logarithmisch geteilt sind.

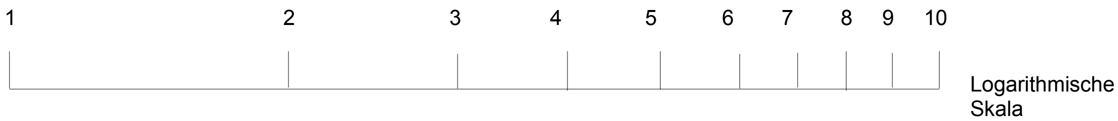

Abb.6

Auffällig an dieser Skala ist, dass der linke Teil gedehnt und der rechte gestaucht ist. Aus diesem Grunde wird eine linksgipfelige Kurve mehr oder weniger gut symmetrisiert, denn die Abstände der links von y_{max} eng beieinander liegenden Werte werden etwas gestreckt und die rechts davon etwas gestaucht. Um Zwischenwerte eintragen zu können, müssen in entsprechender Weise die Logarithmen der Zwischenwerte ermittelt werden. Wie das geschieht, das zeigen wir hier für den Bereich 1 bis 2 der Abb.6. Wie wir vorgehen, wenn an der logarithmischen Skala z.B. die Werte 1,1; 1,2; 1,3 ... 1,9 eingetragen werden sollen, das zeigt die Abb. 7. Um die Zahlen 1,1 bis 1,9 besser positionieren zu können, haben wir die Strecke 1 bis 3 in Abb.7 etwas gespreizt. Wir tragen in Abb.7 die Zwischenwerte als Hilfsskala metrisch ein. Also zwischen 0 und 3 die Werte 0,01; 0,02; 0,03 ... 0,29. Bei Bedarf auch genauer. Dann ermitteln wir die Logarithmen für 1,1 bis 1,9 (Tabelle 3) und tragen sie in Abb.7 an der Skala ab.

Zahlen	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
Logarithmen*10	0.41	0.79	1.14	1.46	1.76	2.04	2.3	2.55	2.79

Tabelle 3

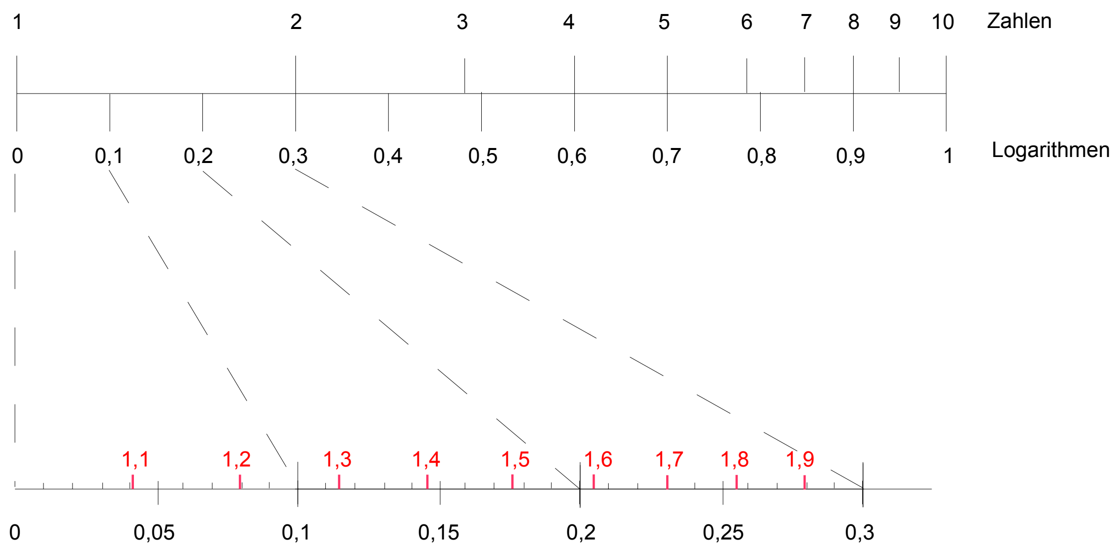

Abb. 7

Entfernen wir nun die Logarithmen (Hilfsskala) unter der Abszisse, dann haben wir in Abb.8 eine logarithmische Skala über den Bereich 0 bis 2,0. Beachten Sie, dass die Abstände nicht metrisch gleich sind!

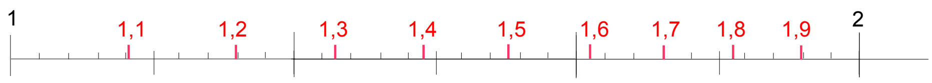

Abb.8

In entsprechen der Weise könnten wir den Rest der Skala unterteilen.

Die Klassenmitten in Beispiel 1 haben eine Spanne von 4,5 bis 11,5. Um sie auf der logarithmischen Abszisse aufzutragen, benötigen wir zwei Zehnerpotenzen, wozu wir zwei der soeben erstellten Skalen hintereinander zeichnen wie das die Abb.9 zeigt.

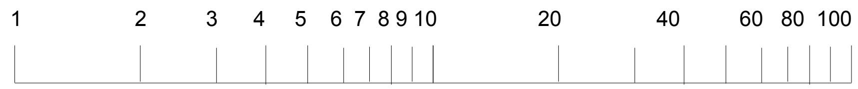

Abb.9

Um die Werte 4,5; 5,0; 5,5; 6,5 bis 11,5 eintragen zu können, müssen die Abstände zwischen zwei Zahlen der Skala wie vorhin gezeigt wieder logarithmisch geteilt werden. Wir dürfen nicht den Fehler begehen, die Abstände metrisch zu teilen. Wir dürfen die 4,5 also nicht in der metrischen Mitte zwischen 4 und 5 eintragen. Wenn Sie sich die Mühe machen und eine Graphik mit der in Abb.9 gezeigten Abszisse erstellen, dann erhalten Sie den gleichen Kurvenverlauf wie in Abb.3.

Damit haben wir jetzt zwar den graphischen Hinweis auf eine logarithmische Normalverteilung, aber noch keine Möglichkeit, die Daten einem Signifikanztest unterziehen zu können. Um normalverteilte Daten zu erhalten müssen wir die Originaldaten (Klassenmitten) logarithmieren.

9.2.3 Logarithmieren der Daten

Zur logarithmischen Transformation müssen wir für die Daten - stellvertretend stehen hier die Klassenmitten - die Logarithmen ermitteln. In Anlehnung an die Transformation bei der Standardisierung der Normalverteilung bezeichnen wir die transformierten Werte mit dem Zeichen z. Bei logarithmischer Transformation gilt also $z = \log x$. Die Logarithmen der Klassenmitten sind in der Tabelle 4 rot eingetragen.

[Würden wir mit nicht klassierten Werten sondern mit den Originalwerten arbeiten, so müsste die gesamte Berechnung mit den Logarithmen der Originaldaten durchgeführt werden.]

KM	4.5	5.0	5.5	6	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5
H	1	2	4	6	9	11	8	6	5	4	3	2	2	1	1
z = log KM	0,653	0,699	0,740	0,778	0,813	0,845	0,875	0,903	0,929	0,954	0,978	1,000	1,021	1,041	1,061

Tabelle 4

Wir tragen die absoluten Häufigkeiten gegen die Logarithmen der Klassenmitten auf einer metrischen Abszisse auf. Die Abb.10 zeigt die gleiche symmetrische Kurve wie Abb.3. Die Abstände der Daten links von y_{max}. sind auseinander gezogen worden und die der Daten rechts davon sind etwas gestaucht. Selbstverständlich interpretieren wir das Ergebnis wie bei Abb.3. Mit den Logarithmen der Daten könnten wir nun einen t-Test durchführen, was einem späteren Kapitel vorbehalten bleibt.

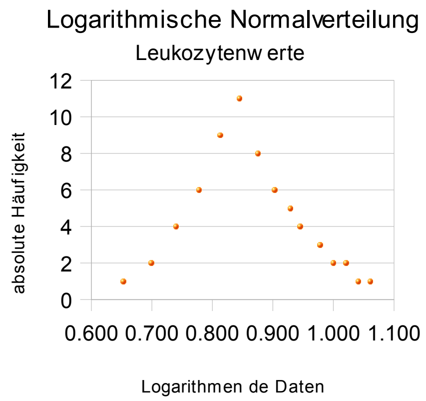

Abb.10

Beide Verfahren, die Darstellung der Daten auf logarithmischer Abszisse und die Darstellung der logarithmierten Daten auf metrischer Abszisse zeigen, dass die Leukozytenzahlen in Zellen/L logarithmisch normalverteilt sind.

Als Grund für die bei vielen biologischen Daten vorkommende „Logarithmische Normalverteilung" wird in der Literatur angegeben, dass die Zufallseinwirkungen, die den Wert einer Variablen beeinflussen, nicht additiv wirken wie bei normalverteilten Daten, sondern oft multiplikativ. Für physiologisch Interessierte sei an das Weber-Fechnersche „Gesetz" erinnert.

Und nun in Kurzfassung das

Beispiel 2 Die folgenden, wiederum fiktiven Daten sollen als Verteilungskurve dargestellt werden. Wenn die Verteilung linksgipfelig ist, dann soll geprüft werden, ob die Daten log-normalverteilt sind.

Messwert in mm	10,0	12,6	15,8	20,0	25,1	31,6	39,8	50,1	63,1	79,1	100,0
absolute Häufigkeit	5	10	20	30	50	70	50	30	20	10	5

Tabelle 5

Nach Bearbeitung der Daten mit einem Tabellenkalkulationssystem erhalten wir die linksgipfelige Verteilung in Abb.11.

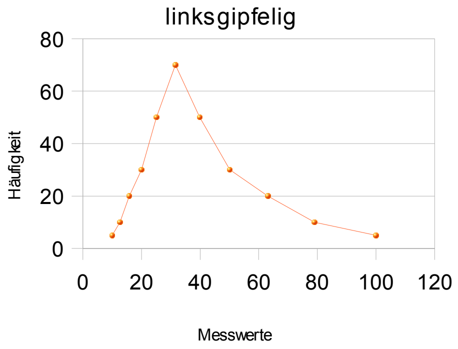

Abb.11

Wir transformieren also die Messwerte logarithmisch und erhalten Tabelle 6.

z = log Messwert in mm	1,000	1,100	1,199	1,301	1,400	1,500	1,600	1,700	1,800	1,900	2,000
absolute Häufigkeit	5	10	20	30	50	70	50	30	20	10	5

Tabelle 6

Da wir mit diesen Werten eine symmetrische Kurve erhalten, gehen wir davon aus, dass die Daten Logarithmisch normalverteilt sind.

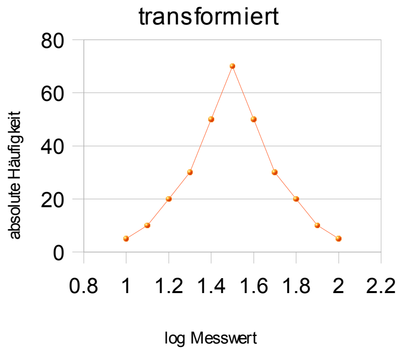

Abb.12

Eine Transformation empirisch gewonnener Daten führt nun nicht unbedingt zu dem gewünschten Erfolg wie bei unseren fiktiven Werten. Erscheinen uns die logarithmisch transformierten Daten einer Normalverteilung nicht hinreichend approximiert, so können wir versuchen, mit anderen Transformationen zum Erfolg zu kommen. Dabei muss ein wenig experimentiert werden um den optimalen Algorithmus zu finden. Versuchen Sie mal einen selbst erstellten, fiktiven linksgipfeligen Datensatz doppeltlogarithmisch zu transformieren (z = log (log x) oder mit einem reziproken Faktor (z = 1/x). Je nach Datenstruktur erhalten Sie damit auch eine approximierte Symmetrisierung der Kurve.

9.3 Transformation rechtsgipfliger Verteilungen

Rechtsgipfelige Verteilungen kommen in Biologie und Medizin selten vor. Die entsprechende Transformation muss mit dem Ziel durchgeführt werden, dass der steile rechte Schenkel der Verteilung gestreckt wird und der linke gestaucht. Umkehrungen der eben genannten Verfahren können hier zum Ziel führen. Z.B. eine Potenztransformation mit $z = x^a$ wobei wir mit dem Wert für den Exponenten a experimentieren müssen. In der Literatur wird auch eine Reziproke Wurzeltransformation mit $z = \sqrt[2]{(1/x)}$ angegeben, wobei hier der Wurzelexponent so variiert werden muss, bis gegebenenfalls ein verwertbares Ergebnis herauskommt. Auch hier könnten Sie mit selbst erstellten fiktiven Daten experimentieren. Die Daten können aber so verteilt sein, dass wir nicht zu einer Normalverteilung kommen.