Kantonsschule XY
Abteilung Mathematik
Strasse 99
9999 Güggliälp

HS 2008

Sonderwoche

Deskriptive Statistik

Simon-Lukas.Rinderknecht@gmx.ch

25. Mai 2009

Inhaltsverzeichnis

1 Einführung

1.1 Etymologie (Herkunft des Wortes *Statistik*)

Das Wort Statistik stammt vom lateinischen „*statisticum*“ („den Staat betreffend“). Die deutsche Statistik, eingeführt von Gottfried Achenwall (1749), bezeichnete ursprünglich die Lehre von den Daten über den Staat, also Staatstheorie. Im 19. Jahrhundert hatte der Engländer Sir John Sinclair das Wort erstmals in seiner heutigen Bedeutung des allgemeinen Sammelns und Auswertens von Daten benutzt.

1.2 Ein überblick

Von Statistiken wird gefordert, dass sie „objektiv“ (unabhängig vom Standpunkt des Statistikerstellers), „reliabel“ (verlässlich), „valide“ (überkontextuell gültig), „signifikant“ (bedeutend) und „relevant“ (wichtig) sind.

Die Statistik wird in die folgenden drei Teilbereiche eingeteilt:

Deskriptive Statistik: Die *deskriptive* Statistik (auch beschreibende Statistik oder empirische Statistik): mit der vorliegende Daten in geeigneter Weise beschrieben und zusammengefasst werden. Mit ihren Methoden verdichtet man quantitative Daten zu Tabellen, graphischen Darstellungen und Kennzahlen. Bei einigen Institutionen, z. B. dem Bundesamt für Statistik in Neuchâtel (BFU), ist die Erstellung solcher Statistiken die Hauptaufgabe.

Induktive Statistik: Die *induktive* Statistik (auch mathematische Statistik, schließende Statistik oder Inferenzstatistik): In der induktiven Statistik leitet man aus den Daten einer Stichprobe Eigenschaften einer Grundgesamtheit ab. Die Wahrscheinlichkeitstheorie liefert die Grundlagen für die erforderlichen Schätz- und Testverfahren.

Explorative Statistik: Die *explorative* Statistik (hypothesen-generierende Statistik, Datenschürfung (data mining)): Methodisch eine Zwischenform der beiden vorgenannten Teilbereiche, bekommt als Anwendungsform jedoch zunehmend eine eigenständige Bedeutung. Mittels deskriptiver Verfahren und induktiver Test-Methoden sucht sie systematisch mögliche Zusammenhänge (oder Unterschiede) zwischen Daten in vorhandenen Datenbeständen und will sie zugleich in ihrer Stärke und Ergebnissicherheit bewerten. Die so gefundenen Ergebnisse lassen sich als Hypothesen verstehen, die erst, nachdem darauf aufbauende, induktive Testverfahren mit entsprechenden (prospektiven) Versuchsplanungen sie bestätigten, als statistisch gesichert gelten können.

Gemeinsam mit der Wahrscheinlichkeitstheorie definiert die mathematische Statistik das mathematische Teilgebiet der Stochastik.

1.3 NZZ-Folio Leitartikel: Wieso haben reiche Männer wenig Haare auf dem Kopf? Wie viele Schweizer gehen zur Kirche? Was ist das geometrische Mittel?

Aus NZZ Folio 01/06

1.3.1 Mittelwert und Streuung

Ein Schweizer isst zwei Dutzend Cervelas im Jahr. Natürlich nur im Durchschnitt über alle Bürger des Landes, und damit sind wir schon mitten in der Statistik: Viele Schweizer essen überhaupt keine Cervelas, andere können nie genug davon bekommen. In der Summe ergibt das 160 Millionen. Diese Summe, geteilt durch die Zahl der Summanden, heisst auch arithmetisches Mittel; das ist für viele der Inbegriff von Durchschnitt überhaupt.

Für die meisten Zwecke reicht diese Art von Durchschnitt völlig aus. Aber zuweilen kann das arithmetische Mittel auch in die Irre führen: Angenommen, wir geben einem Vermögensverwalter 100 000 Franken. Nach einem Jahr werden daraus 160 000 Franken – ein Plus von 60 Prozent. Das Jahr darauf fällt unser Vermögen auf 80 000 Franken – ein Minus von 50 Prozent. Das arithmetische Mittel der beiden Renditen von einmal +60 und einmal –50 Prozent ist (+60 – 50) : 2 = + 5 Prozent. Anders gesagt: Wir haben am Ende weniger als am Anfang, aber im Durchschnitt nimmt der Wert unseres Vermögens in jeder Periode zu!

Profis wissen natürlich, dass man Wachstumsraten niemals arithmetisch mitteln darf. Der korrekte Durchschnitt ist hier jene jährliche Rendite, die in zwei Jahren aus 100 000 Franken 80 000 Franken macht, das sind (leicht gerundet) – 10,55 Prozent: Nach einem Jahr werden so aus den anfänglichen 100 000 Franken damit 10,55 Prozent weniger, das sind 89 450 Franken, das nächste Jahr werden aus diesen 89 450 Franken nochmals 10,55 Prozent weniger, das sind dann (bis auf Rundungsfehler) 80 000 Franken. Diese Durchschnittsrendite von – 10,55 Prozent findet man über das geometrische Mittel der beiden sogenannten Wachstumsfaktoren 1,6 und 0,5. Es wird errechnet, indem man die Wurzel aus 1,6 x 0,5 = 0,8 zieht, das ergibt 0,8945. Von diesem geometrischen Mittel der Wachstumsfaktoren ist dann noch 1 abzuziehen: 0,8945 – 1 = – 0, 1 055.

Regelmässige Proteste ruft das arithmetische Mittel bei Meldungen der Art hervor, dass etwa niedergelassene ärzte in der Schweiz im Jahr im Durchschnitt 205 000 Franken Einkommen erzielen. «Stimmt überhaupt nicht, viel zu hoch!» hört man dann ärztefunktionäre klagen. «Drei Viertel aller ärzte verdienen weniger, der Durchschnitt beträgt nur 165 000 Franken!» Das ist auch so, nur haben diese Kritiker nochmals einen anderen Durchschnitt im Sinn, den

sogenannten Zentralwert oder Median. Der Median ist der Wert, der in der Mitte steht, wenn man alle Einkommen der Grösse nach sortiert. Bei drei Einkommen 1,3 und 8 ist der Zentralwert 3, das arithmetische Mittel aber grösser, nämlich 4, und das ist typisch für Merkmale wie Einkommen, Vermögen oder Grundbesitz, wo oft einige wenige sehr viel mehr haben als alle anderen. Hier liegt der Median in aller Regel unter dem arithmetischen Mittel; er ist unempfindlich gegen hohe Werte am rechten Rand, die wie ein Magnet das arithmetische Mittel nach oben ziehen. Statistiker sagen dazu auch «robust».

Und oft blenden natürlich Durchschnitte wichtige Informationen einfach aus: Wenn ich im Durchschnitt jeden Tag des Monats einen Viertel Rotwein trinke, aber alle am gleichen Tag, bekomme ich eine Alkoholvergiftung und bin tot. Trinke ich dagegen jeden Tag nur einen, lebe ich sogar länger als Leute, die nie Rotwein trinken. Der Durchschnitt ist in beiden Fällen gleich, aber die Abweichung vom Durchschnitt ist im ersten Fall erheblich grösser. Deshalb fügt man Durchschnitten am besten immer auch ein Mass für die Abweichung vom Durchschnitt bei, wie die Schwankungsbreite oder die Standardabweichung.

Die Schwankungsbreite ist einfach der Abstand zwischen dem grössten und dem kleinsten Wert; die Standardabweichung ist die «durchschnittliche» Abweichung vom Durchschnitt. Bei sogenannten normalverteilten Daten liegen 95 Prozent der Fälle weniger als zwei Standardabweichungen vom arithmetischen Mittel entfernt. Wenn man Sätze hört wie «Ein erwachsener Mitteleuropäer hat einen IQ von 100 + / – 15», so ist in aller Regel das damit gemeint. «Normalverteilt» soll dabei heissen, dass sich sehr vieles – früher glaubte man sogar: fast alles –, was sich auf dieser Erde messen oder wiegen lässt, auf eine ganz bestimmte Weise um den Durchschnitt streut: Die Masse drängt sich dicht darum herum, aber mit wachsender Entfernung nimmt die Häufigkeit der Werte dann glockenförmig ab.

1.3.2 Korrelation und Kausalität

Oft ist bei den Objekten einer Untersuchung mehr als nur eine einzige Variable von Interesse: bei Immobilien die Lage, die Grösse und der Preis; bei Partnerschaftsinseraten in der NZZ das Geschlecht, das Alter, die Körpergrösse, der Beruf; bei Patienten in der Klinik der Blutdruck und die Dosis eines blutdrucksenkenden Medikaments. Da wüsste man oft gerne: hängen diese Variablen zusammen – und wenn ja, wie? Das führt in den Bereich der modernen Statistik, der sich mit Abhängigkeiten – sogenannten Korrelationen – und Kausalbeziehungen befasst.

Die Grafik «Korrelation und Kausalität» stellt die Lebenserwartung der Männer und das durchschnittliche Pro-Kopf-Einkommen in 25 Schweizer Kantonen einander gegenüber. Zusätzlich sind auch noch die beiden arithmetischen Mittelwerte eingetragen. Basel-Stadt ist nicht dabei, weil die Menschen dort zwar viel verdienen (im Durchschnitt 99 000 Franken jährlich, das ist Landesrekord), aber dennoch früher sterben als in den meisten anderen Kantonen. Solche Datenpunkte, die sich von allen anderen drastisch unterscheiden, heissen Aus-

reisser; die behandelt man besser getrennt (wobei wir das Spekulieren über diesen Ausreisser hier den Soziologen und Demographen überlassen wollen).

Im Grossen und Ganzen leben Männer in Kantonen mit hohem Durchschnittseinkommen länger (Punkte in der Grafik oben rechts); Männer in Kantonen mit tiefem Durchschnittseinkommen sterben früher (Punkte in der Grafik unten links). Dies ist ein Beispiel für eine positive Korrelation: je mehr vom einen, desto mehr auch vom andern. Eine negative Korrelation dagegen bedeutet: je mehr vom einen, desto weniger vom andern. (Je mehr Regenschirme verkauft werden, desto weniger Sonnencrème wird abgesetzt.) Das Mass der Abhängigkeit von zwei Variablen (Durchschnittseinkommen / Lebenserwartung) ist der sogenannte Korrelationskoeffizient. Er liegt zwischen minus eins (maximale negative Korrelation) und plus eins (maximale positive Korrelation) und drückt aus, wie sicher man zum Beispiel vom Durchschnittseinkommen auf die Lebenserwartung schliessen kann. In unserem Beispiel hat der Korrelationskoeffizient den Wert 0,49.

Oft wird aus einer positiven oder negativen Korrelation auf eine positive oder negative Kausalbeziehung geschlossen. Das ist nicht immer richtig. Es gibt zum Beispiel bei erwachsenen Männern eine bemerkenswerte negative Korrelation zwischen dem Einkommen und der Zahl der Haare auf dem Kopf. Aber weder sind die Haare für das Einkommen noch ist das Einkommen für die Haare verantwortlich zu machen – diese negative Korrelation kommt dadurch zustande, dass beide Variablen von einer dritten Variablen, dem Lebensalter, abhängen: mit wachsendem Alter nimmt das Einkommen zu, und die Haare fallen aus. Bei der Interpretation von Korrelationen ist also immer darauf zu achten, dass man keine dritte, eigentlich kausale Variable übersieht.

1.3.3 Stichproben und Umfragen

Bevor man Mittelwerte, Standardabweichungen oder Korrelationskoeffizienten ausrechnet, muss man die Daten natürlich erst einmal haben. Dafür behilft man sich oft mit Stichproben; sie reichen für viele Zwecke völlig aus. Wie bei einer Polizeikontrolle, wo man aus einer winzigen Stichprobe unseres Blutes den gesamten Alkoholanteil problemlos abliest, lässt sich auch aus einer Stichprobe von 1000 oder 2000 befragten Bürgern recht präzise hochrechnen, wie viele Schweizer insgesamt einen EU-Beitritt ablehnen oder sonntags in die Kirche gehen.

Vorausgesetzt, die Grundgesamtheit, aus der die Stichprobe kommt, wird wie ein Kartenspiel oder der Inhalt einer Urne vorher gut gemischt. Unser Blut besorgt dieses Mischen mit Hilfe physikalisch-chemischer Gesetze von allein. Bei der Bevölkerung der Schweiz wird das Durchmischen nur simuliert. Das Standardverfahren dafür ist eine sogenannte einfache Zufallsstichprobe: alle Schweizer haben die gleiche Chance, in die Stichprobe zu kommen. Wir verteilen quasi Nummern, für jeden erwachsenen Schweizer Bürger eine, notieren diese Nummer auf einer Lottokugel, legen die Kugeln in eine grosse Urne, schütteln kräftig und ziehen

tausend Kugeln zufällig heraus. Aufgrund dieser Stichprobe wissen wir mit grosser Zuverlässigkeit, welcher Anteil der erwachsenen Schweizer nicht in die EU will oder sonntags in die Kirche geht.

In der Praxis kann man natürlich nur versuchen, diesem Ideal des Ziehens aus einer Urne möglichst nahe zu kommen. Die Repräsentativität der Stichprobe lässt sich verbessern, wenn man getrennte Urnen für Männer und Frauen oder für die Bürger verschiedener Kantone vorsieht. Solche «geschichteten» Stichproben garantieren, dass die landesweiten Geschlechter- oder kantonalen Proportionen in der Stichprobe erhalten bleiben. Bei einfachen Zufallsstichproben ist das nicht notwendig der Fall.

Abweichungen von diesem zentralen Zufallsprinzip führen zu verzerrten Stichproben – mit zuweilen desaströsen Folgen. Man stelle sich vor, wir fragten zum Sonntagskirchgang nur die Teilnehmer des Hochamts in der Zürcher Liebfrauenkirche. Dann würden hochgerechnet vielleicht 90 Prozent aller Schweizer regelmässig sonntags in die Kirche gehen. Die bisher grösste derartige Pleite widerfuhr der amerikanischen Wochenzeitschrift «Literary Digest» im Jahr 1936: Sie hatte vor der Präsidentenwahl mehrere Millionen US-Bürger befragt (eine nach heutigen Massstäben gewaltige Stichprobe), wen sie zu wählen gedächten. Es siegte mit grossem Vorsprung der Republikaner Landon. Die Wahl gewann jedoch Roosevelt mit über 60 Prozent der Stimmen. Warum die Fehlprognose? Die Stichprobe war aus Telefonregistern und Fahrzeugzulassungen gezogen worden – die meisten Wähler Roosevelts hatten damals aber weder ein Auto noch ein Telefon.

Zusätzliche Verzerrungen drohen ferner immer dann, wenn man die gewünschten Informationen nicht wie die Körpergrösse oder den Stromverbrauch einfach misst oder abliest, sondern erfragt. Aus den USA weiss man, dass mündliche Interviews zu den Themen Abtreibung, Todesstrafe oder Sozialhilfe andere Ergebnisse haben, je nachdem, ob der Interviewer ein Schwarzer oder ein Weisser ist. Auch die Reihenfolge der Fragen und natürlich die konkrete Formulierung sind für das Ergebnis von erheblicher Bedeutung. So fragte etwa die Forscherin Elisabeth Noelle-Neumann einmal eine repräsentative Stichprobe von Arbeitern: «Finden Sie, dass in einem Betrieb alle Arbeiter in der Gewerkschaft sein sollten?» Resultat: dafür 44 Prozent; dagegen 20 Prozent; unentschieden 36 Prozent.

Dann legte sie einer anderen, gleich grossen und ebenfalls repräsentativen Stichprobe die gleiche Frage vor, nur mit der Ergänzung «. . . oder muss man es jedem einzelnen überlassen, ob er in der Gewerkschaft sein will oder nicht?». Ergebnis: dafür 24 Prozent; selbst überlassen 70 Prozent; unentschieden 6 Prozent. Der scheinbar unschuldige Zusatz halbiert die Anhängerschaft der Gewerkschaften von 44 auf nur noch 24 Prozent; zugleich lässt er die Gegner von 20 auf 70 Prozent anwachsen – eine mehr als dreifach grössere Opposition nur wegen eines kleinen Nebensatzes.

Einen grossen Unterschied macht es auch, ob man etwas «verbieten» oder «nicht erlauben» soll. 54 Prozent der Befragten in einer amerikanischen Umfrage meinten, dass die USA öffentliche Angriffe auf die Demokratie verbieten sollten. Erheblich mehr, nämlich 75 Prozent, waren der Meinung, die USA sollten öffentliche Angriffe auf die Demokratie nicht erlauben.

Diese Abhängigkeit der Ergebnisse von der Art der Fragestellung lädt natürlich zur bewussten Irreführung ein. Nach einer Umfrage einer deutschen Gewerkschaft lehnen 95 Prozent der bundesdeutschen Arbeitnehmer das Arbeiten am Samstag ab. Nach einer zeitgleichen Umfrage eines eher unternehmernahen Instituts dagegen sind 72 Prozent aller Arbeitnehmer auch zum Arbeiten am Wochenende bereit. Der Widerspruch erklärt sich durch die jeweiligen Fragebogen. «Votum für das freie Wochenende» steht bei der Gewerkschaft in grossen Lettern obenan. Es folgt eine lange Erläuterung der Mühen, die das Durchsetzen der 5-Tage-Woche die Gewerkschaften gekostet habe, und eine Aufzählung aller Vorteile, die der freie Samstag für die Familie, die Gesellschaft, den Frieden und die Menschheitszukunft bringe, die dann zu der eigentlichen Frage überleitet: «Was entspricht Deiner/Ihrer Meinung? (I) Nach meiner Ansicht wäre die Abschaffung des freien Wochenendes ein schwerer Schlag für Familie, Freundschaften, Partnerschaften, für Geselligkeit, Vereine, den Sport und das Kulturleben; (II) Ich halte den gemeinsamen Freizeitraum des Wochenendes für nicht so wichtig; (III) Weiss nicht / keine Angabe.» Dass hier fast alle wie gewünscht die erste Antwort wählen, sollte niemanden erstaunen.

Genauso suggestiv, wenn auch mit umgekehrter Absicht, fragte das Unternehmerinstitut. Auf die Frage: «Inwieweit wären Sie bereit, samstags zu arbeiten, wenn es für die wirtschaftliche Situation Ihres Unternehmens gut wäre?» bietet es folgende Auswahlmöglichkeiten an: (I) gelegentlich, wenn dafür an einem anderen Tag arbeitsfrei ist; (II) häufiger, wenn dafür ein Zusatzurlaub herauskommt; (III) abwechselnd und (IV) nicht bereit. Auch hier waren die wenigen Kreuze bei «nicht bereit» schon im Fragebogen und in der Art der Fragen angelegt. Solche Umfragen, ob von einem Automobilclub zum Thema Tempolimit, ob von Greenpeace zum Atomausstieg oder von der katholischen Kirche zur Frage der Abtreibung, belügen uns in aller Regel über die wahre Meinung der befragten Menschen.

1.3.4 Was heisst eigentlich «signifikant»?

Zurück zu unserem Ausgangspunkt, dem Cervelas. Im Sommer 2005 liess NZZ-Folio zehn Sorten dieser Wurst von vier Prüfern auf einer Skala von 1 bis 20 benoten; das Ergebnis war im Heft 7/2005 zu lesen. Der beste Cervelas erreichte im Durchschnitt über alle Prüfer 15,5 Punkte, der schlechteste 12,75. Aber ist der am schlechtesten bewertete Cervelas auch wirklich schlechter? Oder können solche Unterschiede auch zufällig zustande kommen? Schliesslich schmeckt ja auch ein und derselbe Cervelas nicht jedem Prüfer immer gleich, seine Bewertungen weichen zufällig, aufgrund der Reihenfolge der Verkostung, aufgrund von unterschiedlichem Appetit und Dutzenden weiterer Faktoren, von der für ihn «wahren» Note mehr oder

weniger nach oben und nach unten ab.

Nehmen wir also einmal an, alle Cervelas wären von der gleichen Qualität; jeder Prüfer hätte für diese Qualität seine eigene «wahre» Bewertung, der Prüfer A zum Beispiel 13,5; diese Note wäre bei Prüfer A für alle Würste gleich und nur durch eine Zufallskomponente überlagert. Dann lässt sich mit einigen Rechenregeln zu Wahrscheinlichkeiten zeigen, dass in der Tat das in NZZ-Folio 7/2005 gemeldete Ergebnis auch durch reinen Zufall erklärt werden könnte. Oder in der Sprache der Statistik: Die beobachteten Unterschiede sind nicht signifikant.

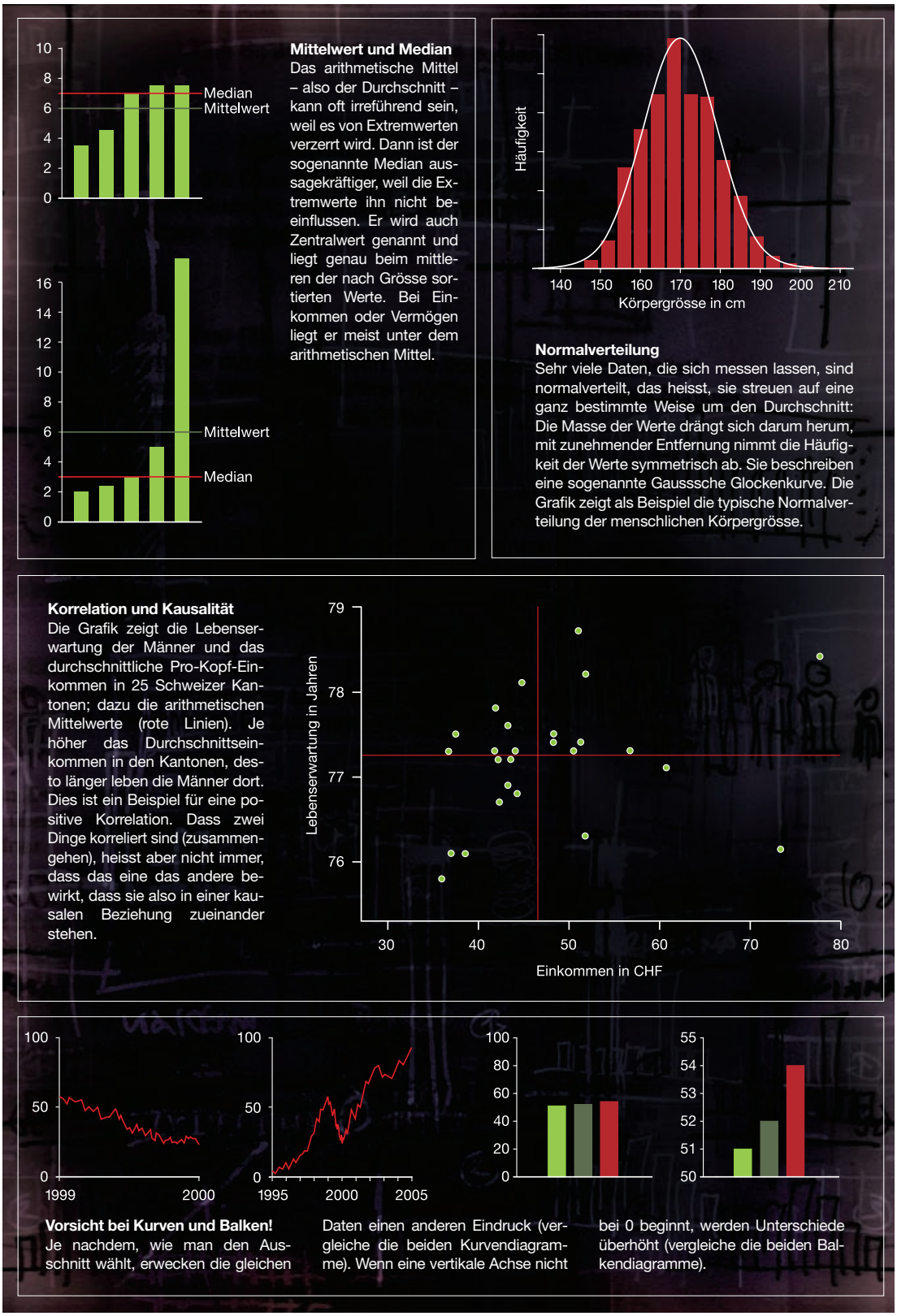
Signifikant dagegen heisst: Ein in den Daten sichtbares Muster ist nur schwer durch Zufall zu erklären. Also, so der Umkehrschluss, steckt ein System dahinter. «Nur schwer durch Zufall zu erklären» meint dabei im Allgemeinen: Wenn wirklich nur der Zufall wirken würde, hätte das beobachtete Muster eine Wahrscheinlichkeit von höchstens 5 Prozent. (Diese Grenze, auch Signifikanzniveau genannt, ist natürlich willkürlich, wenn auch in den meisten Wissenschaften üblich. Mit dem gleichen Recht könnte man auch 1 Prozent oder 10 Prozent verwenden.)

Diese heute in allen Wissenschaften übliche Methode zur Trennung von Zufall und System hat aber einen grossen und auch von den Wissenschaftern gern übersehenen Pferdefuss: Die statistischen Verfahren zur Trennung von Zufall und System zeigen selbst bei Abwesenheit jedes systematischen Einflusses in immerhin 5 Prozent der Fälle dennoch eine Signifikanz an – so sind die Verfahren ja gerade konstruiert. Auch wissenschaftliche Fachzeitschriften und ihre Herausgeber vergessen das nur allzu gerne.

Und so können wir dann in den Medien lesen, dass neun Monate nach einem Stromausfall in X dort die Geburten angestiegen sind, dass Katholiken dümmer sind als Protestanten, dass Knoblauchesser länger leben, dass Manager lieber Fluggesellschaft A als B benutzen, dass die Todesstrafe abschreckt, dass die Todesstrafe nicht abschreckt (je nach Weltanschauung), dass Schwarze krimineller sind als Weisse, dass Chemiefabriken (Starkstromleitungen, Müll deponien) Leukämie erzeugen. Selbstverständlich alles wissenschaftlich abgesichert und hoch signifikant.

Wir lesen jedoch nicht, wie viele andere Studien und Stichproben ohne signifikante Resultate es ausserdem gegeben hat. Wir lesen nicht, in wie vielen Studien Katholiken genauso klug sind wie Protestanten oder Manager lieber Fluglinie B als Linie A benutzen oder Industriebetriebe keine Leukämie erzeugen. Und ehe wir das nicht wissen, lässt sich auch die wahre Bedeutung der angeblich so signifikanten Resultate nicht ermessen.

Walter Krämer ist Professor für Wirtschafts- und Sozialstatistik an der Universität Dortmund. Zu seinen populären Büchern gehören «So lügt man mit Statistik» und «Statistik verstehen: eine Gebrauchsanweisung» (beide als Taschenbuch bei Piper).



Mittelwert und Median
Das arithmetische Mittel – also der Durchschnitt – kann oft irreführend sein, weil es von Extremwerten verzerrt wird. Dann ist der sogenannte Median aussagekräftiger, weil die Extremwerte ihn nicht beeinflussen. Er wird auch Zentralwert genannt und liegt genau beim mittleren der nach Grösse sortierten Werte. Bei Einkommen oder Vermögen liegt er meist unter dem arithmetischen Mittel.

Normalverteilung
Sehr viele Daten, die sich messen lassen, sind normalverteilt, das heisst, sie streuen auf eine ganz bestimmte Weise um den Durchschnitt: Die Masse der Werte drängt sich darum herum, mit zunehmender Entfernung nimmt die Häufigkeit der Werte symmetrisch ab. Sie beschreiben eine sogenannte Gausssche Glockenkurve. Die Grafik zeigt als Beispiel die typische Normalverteilung der menschlichen Körpergrösse.

Korrelation und Kausalität
Die Grafik zeigt die Lebenserwartung der Männer und das durchschnittliche Pro-Kopf-Einkommen in 25 Schweizer Kantonen; dazu die arithmetischen Mittelwerte (rote Linien). Je höher das Durchschnittseinkommen in den Kantonen, desto länger leben die Männer dort. Dies ist ein Beispiel für eine positive Korrelation. Dass zwei Dinge korreliert sind (zusammengehen), heisst aber nicht immer, dass das eine das andere bewirkt, dass sie also in einer kausalen Beziehung zueinander stehen.

Vorsicht bei Kurven und Balken!
Je nachdem, wie man den Ausschnitt wählt, erwecken die gleichen Daten einen anderen Eindruck (vergleiche die beiden Kurvendiagramme). Wenn eine vertikale Achse nicht bei 0 beginnt, werden Unterschiede überhöht (vergleiche die beiden Balkendiagramme).

2 Mittelwerte

Mittelwerte treten in der Mathematik und insbesondere in der Statistik in inhaltlich unterschiedlichen Kontexten auf. In der Statistik ist ein Mittelwert ein sog. *Lageparameter*, also ein aggregierender Parameter einer Verteilung, einer Stichprobe oder Grundgesamtheit. Ziel solcher aggregierender Parameter ist es, die wesentliche Information in einer längeren Reihe von (z. B.) Messdaten in wenigen Daten zu konzentrieren. In der Mathematik treten Mittelwerte, insbesondere die drei klassischen Mittelwerte (Arithmetisches, Geometrisches und Harmonisches Mittel) bereits in der Antike auf. Pappos von Alexandria kennzeichnet 10 verschiedene Mittelwerte m von 2 Zahlen a und b ($a < b$) durch spezielle Werte des Streckenverhältnisses $(b - m)/(m - a)$. Auch die Ungleichung zwischen harmonischem, geometrischem und arithmetischem Mittel ist in der Antike bereits bekannt und geometrisch interpretiert.

Spezifisch für nimmersättliche Alleswisser: Im 19. und 20. Jahrhundert spielen Mittelwerte in der Analysis eine spezielle Rolle, dort im wesentlichen im Zusammenhang mit berühmten Ungleichungen und wichtigen Funktionseigenschaften wie Konvexität (Hölder-Ungleichung, Minkowski-Ungleichung, Jensensche Ungleichung usw.). Dabei wurden die klassischen Mittelwerte in mehreren Schritten verallgemeinert, zunächst zu den Potenzmittelwerten und diese wiederum zu den quasi-arithmetischen Mittelwerten. Die klassische Ungleichung zwischen harmonischem, geometrischem und arithmetischem Mittel geht dabei über in allgemeinere Ungleichungen zwischen Potenzmittelwerten bzw. quasi-arithmetischen Mittelwerten.

Im Folgenden seien $x_1, ..., x_n$ gegebene reelle Zahlen, in der Statistik etwa Messwerte, deren Mittelwert berechnet werden soll.

2.1 Arithmetisches Mittel

2.1.1 Definition

Das arithmetische Mittel (auch Durchschnitt) ist ein rechnerisch bestimmter Mittelwert. Es ist so definiert:

$$\bar{x}_{arithm} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ... + x_n}{n} \tag{2.1}$$

2.1.2 Anwendungsbeispiel

Ein Auto fährt eine Stunde lang 100 km/h und die darauf folgende Stunde 200 km/h. Mit welcher konstanten Geschwindigkeit muss ein anderes Auto fahren, um denselben Weg ebenfalls in 2 Stunden zurückzulegen? (Rechne - wer kann! Antwort: 15 km/h)

2.1.3 Spezialfall: Gewichtetes arithmetisches Mittel

Das gewichtete Mittel wird beispielsweise verwendet, wenn man Mittelwerte x_i, aus n Stichproben der gleichen Grundgesamtheit mit verschiedenen Stichprobenumfängen wi miteinander kombinieren will:

$$\bar{x}_{gewarithm} = \frac{\sum_{i=1}^{n} \omega_i x_i}{\sum_{i=1}^{n} \omega_i} \tag{2.2}$$

Beispiel: Ein Schüler erarbeitet sich folgende Noten: $\{6, 5, 4, 6, 5.5\}$. Die 4 ist eine Streichnote und die 5 zählt nur halb. Wie gross sind die Gewichtungen der Noten numerisch? Was ist die Schlussnote?

2.2 Geometrisches Mittel

2.2.1 Definition

Das geometrische Mittel ist die n-te Wurzel aus dem Produkt der positiven Zahlen $x_1, ..., x_n$.

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} \tag{2.3}$$

Es ist in der Statistik ein geeignetes Lagemaß für Größen, von denen das Produkt anstelle der Summe interpretierbar ist, z. B. von Verhältnissen oder Wachstumsraten.

2.2.2 Anwendungsbeispiel

Ein Guthaben G wird im ersten Jahr mit zwei Prozent, im zweiten Jahr mit sieben und im dritten Jahr mit fünf Prozent verzinst. Welcher über die drei Jahre konstante Zinssatz p hätte zum Schluss das gleiche Kapital ergeben?

Guthaben G_{Ende} am Ende des dritten Jahres:

$$G_{Ende} = \left(1 + \frac{2}{100}\right)\left(1 + \frac{7}{100}\right)\left(1 + \frac{5}{100}\right)G$$

oder mit Zinsfaktoren geschrieben

$$G_{Ende} = 1.2 \cdot 1.7 \cdot 1.5 \cdot G$$

Mit konstantem Zinssatz p und zugehörigen Zinsfaktor $1 + p$ ergibt sich am Ende ein Guthaben von

$$G_{Ende} = (1 + p)^3 \cdot G$$

Mit $G_{konst} = G_{Ende}$ ergibt sich

$$(1 + p)^3 \cdot G = 1.2 \cdot 1.7 \cdot 1.5 \cdot G$$

und damit berechnet sich der durchschnittliche Zinsfaktor $1 + p$ zu

$$1 + p = \sqrt[3]{1.2 \cdot 1.7 \cdot 1.5} \approx 1.04646$$

Der durchschnittliche Zinssatz beträgt also ca 4.646%. Allgemein berechnet sich der durchschnittliche Zinsfaktor also aus dem geometrischen Mittel der Zinsfaktoren der einzelnen Jahre. Wegen der Ungleichung vom arithmetischen und geometrischen Mittel ist der durchschnittliche Zinssatz *kleiner* oder bestenfalls gleich dem arithmetischen Mittel der Zinssätze, welches in diesem Beispiel $\frac{14}{3}\% \approx 4.667\%$ beträgt.

2.2.3 Spezialfall: Gewichtetes geometrisches Mittel

Analog zum gewichteten arithmetischen Mittel lässt sich ein mit den Gewichten $w_i > 0$ gewichtetes geometrisches Mittel definieren:

$$\bar{x}_{gewgeom} = \sqrt[\omega]{\prod_{i=1}^{n} x_i^{\omega_i}} = \sqrt[\omega]{x_1^{\omega_1} \cdot x_2^{\omega_2} \cdot ... \cdot x_n^{\omega_n}} \qquad \text{wobei } \omega = \sum_{i=1}^{n} \omega_i = 1 \tag{2.4}$$

2.3 Harmonisches Mittel

2.3.1 Definition

Das harmonische Mittel ist definiert als

$$\bar{x}_{harm} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{2.5}$$

Durch Bildung des Kehrwertes erhält man

$$\frac{1}{\bar{x}_{harm}} = \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n}$$

der Kehrwert des harmonischen Mittels ist also das arithmetische Mittel der Kehrwerte.

2.3.2 Anwendungsbeispiel

Beispiel für das harmonische Mittel von 5 und 20:

$$\frac{2}{\frac{1}{5} + \frac{1}{20}} = \frac{2}{\frac{1}{4}} = 8$$

Mit dieser Formel ist das harmonische Mittel zunächst nur für von Null verschiedene Zahlen x_i definiert. Geht aber einer der Werte x_i gegen Null, so existiert der Grenzwert des harmonischen Mittels und ist ebenfalls gleich Null. Daher ist es sinnvoll, das harmonische Mittel als Null zu definieren, wenn mindestens eine der zu mittelnden Größen gleich Null ist.

2.3.3 Spezialfall: Gewichtetes harmonisches Mittel

Auch hier lässt sich ein mit den Gewichten $\omega_i > 0$ gewichtetes harmonisches Mittel definieren:

$$\bar{x}_{gewharmon} = \frac{\sum_{i=1}^{n} \omega_i}{\sum_{i=1}^{n} \frac{\omega_i}{x_i}} \tag{2.6}$$

Fährt man eine Stunde mit 50 km/h und dann eine Stunde mit 100 km/h, so legt man insgesamt 150 km in 2 Stunden zurück; die Durchschnittsgeschwindigkeit ist 75 km/h, also das arithmetische Mittel von 50 und 100. Bezieht man sich hingegen nicht auf die benötigte Zeit, sondern auf die durchfahrene Strecke, so wird die Durchschnittsgeschwindigkeit durch das harmonische Mittel beschrieben: fährt man 100 km mit 50 km/h und dann 100 km mit 100 km/h, so legt man 200 km in 3 Stunden zurück, die Durchschnittsgeschwindigkeit ist 66 2/3 km/h, also das harmonische Mittel von 50 und 100.

Allgemein gilt: Benötigt man für die Teilstrecke s_1 die Zeit t_1 (also Durchschnittsgeschwindigkeit $v_1 = s_1/t_1$) und für die Teilstrecke s_2 die Zeit t_2 (also Durchschnittsgeschwindigkeit $v_2 = s_2/t_2$, so gilt für die Durchschnittsgeschwindigeit über die gesamte Strecke

$$v = \frac{s_1 + s_2}{t_1 + t_2} = \frac{s_1 + s_2}{\frac{s_1}{v_1} + \frac{s_2}{v_2}} = \frac{t_1 v_1 + t_2 v_2}{t_1 + t_2}$$

Die Durchschnittsgeschwindigkeit ist also das mit den Wegstrecken gewichtete harmonische Mittel der Teilgeschwindigkeiten oder das mit der benötigten Zeit gewichtete arithmetische Mittel der Teilgeschwindigkeiten.

3 Streuungswerte

3.1 Schwankungsbreite

3.1.1 Definition

Die Schwankungsbreite ist das Intervall zwischen dem grössten und dem kleinsten Wert der Verteilung. Die Schwankungsbreite ist ein grober Schätzwert für die Streuung der Verteilung.

$$s = x_{max} - x_{min} \tag{3.1}$$

3.2 Standardabweichung

Die Standardabweichung ist ein Maß, das beschreibt, wie sehr ein Sachverhalt „streut“. Sie wird berechnet, indem man die Abstände der Messwerte vom Mittelwert quadriert, addiert und durch die Anzahl der Messwerte teilt.

3.2.1 Definition

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \tag{3.2}$$

3.2.2 Anwendungsbeispiel

Die Bestimmung der Gewichte von 5 Tabletten ergab in Gramm: $\{0.62; 0.64; 0.68; 0.65\}$
Berechne die Standardabweichung der Verteilung:

s =

TI-92 Voyager

Die Standardabweichung einer Werteliste `liste` lässt sich mit dem Taschenrechner direkt berechnen. Der Befehl lautet: `stddev(liste)`.
Aufgabe überprüfe mit dem Taschenrechner die berechnete Standardabweichung.

4 Stichproben / Sampling

4.1 Definition

Als Stichprobe bezeichnet man eine Teilmenge einer Grundgesamtheit[1], die unter bestimmten Gesichtspunkten ausgewählt wurde. Mit Stichproben wird in Anwendungen der Statistik (etwa in der Marktforschung, aber auch in der Qualitätskontrolle und in der naturwissenschaftlichen, medizinischen und psychologischen Forschung) häufig gearbeitet, da es oft nicht möglich ist, die Grundgesamtheit, etwa die Gesamtbevölkerung oder alle hergestellten Exemplare eines Produkts, zu untersuchen. Grundgedanke der Zuhilfenahme von Stichproben ist das Induktionsprinzip, bei dem von besonderen auf allgemeine Fälle geschlossen wird.

Um die einzelnen Elemente einer Stichprobe zu erhalten, stehen verschiedene Auswahlverfahren zur Verfügung. Die korrekte Wahl des Auswahlverfahrens ist wichtig, da die Stichprobe repräsentativ sein muss, um auf die Grundgesamtheit schließen zu können (siehe dazu z. B. Hochrechnung). Entscheidend ist eine vernünftige Probenahme, die über den Erfolg der Aussage entscheidet. Häufig sind mehrere Tests notwendig um sicherzustellen, dass tatsächlich rational entschieden wurde.

4.2 Stichproben-Typen

Sollen zwei Stichproben mittels statistischer Tests miteinander verglichen werden, so muss zwischen abhängigen und unabhängigen Stichproben unterschieden werden:

Abhängige Stichproben: Elemente von zwei (oder mehr) Stichproben können einander jeweils paarweise zugeordnet werden. Beispiel: Stichprobe 1 besteht aus Personen vor der Behandlung mit einem bestimmten Medikament, und soll verglichen werden mit Stichprobe 2, welche aus den gleichen Personen nach der Behandlung besteht.

Unabhängige Stichproben: Es besteht kein Zusammenhang zwischen den Elementen der Stichproben. Dies ist beispielsweise der Fall, wenn die Elemente der Stichproben jeweils aus unterschiedlichen Population kommen (z. B. Stichprobe 1 besteht aus Frauen, Stichprobe 2 aus Männern), oder wenn Personen nach dem Zufallsprinzip in zwei oder mehrere Gruppen aufgeteilt werden.

[1] In der empirischen Forschung bezeichnet die Grundgesamtheit (auch Population) die Menge aller potentiellen Untersuchungsobjekte für eine bestimmte Fragestellung. Aus pragmatischen Erwägungen wird normalerweise nicht die Grundgesamtheit, sondern eine Stichprobe untersucht, die für die Grundgesamtheit repräsentativ ist.

4.3 Auswahlverfahren

Ein Auswahlverfahren ist die Art und Weise, wie Personen oder Dinge für einen Zweck ausgewählt werden. Die Statistik beschäftigt sich in der Kombinatorik mit grundsätzlich möglichen Auswahlen. In der Empirie werden mehrere Verfahren (Stichprobenverfahren) zur Auswahl einer repräsentativen Stichprobe unterschieden. Die unterschiedlichen Wahrscheinlichkeiten eines Elementes der Grundgesamtheit, je nach Auswahlverfahren in die Stichprobe zu gelangen, nennt man Einschlusswahrscheinlichkeit. Als Auswahlverfahren werden auch kommerzielle Verfahren bezeichnet, die an Repräsentativität nicht interessiert sind. Die tatsächliche Auswahl der Auskunftgebenden erfolgt z. B. mit dem Random-Route-Verfahren und dem Schwedenschlüssel.

In der Empirie dient das Auswahlverfahren (auch Stichprobenverfahren) der Ermittlung einer repräsentativen Stichprobe. Man unterscheidet generell Stufung, Schichtung und Klumpung. Die verschiedenen Typen von Auswahlverfahren können folgendermaßen charakterisiert werden:

Zufallsauswahlverfahren: Bei einem Zufallsauswahlverfahren (auch Wahrscheinlichkeitsauswahl, Zufalls-Stichprobe, Zufallsauswahl, Random-Sample) hat jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit (größer Null), in die Stichprobe zu gelangen. Das erfordert die vorherige Erstellung eines Gesamtverzeichnisses aller Elemente der Grundgesamtheit. Man unterscheidet einstufige und mehrstufige Zufallsauswahlverfahren. Nur bei Zufallsauswahlen sind streng genommen die Methoden der induktiven Statistik anwendbar.

Systematische Stichproben: Auswahlverfahren, bei denen subjektive Erwägungen die Auswahl der Zielpersonen bestimmen. Es werden Vorinformationen über die auszuwählenden Fälle genutzt. Verallgemeinerungen sind auf der Basis mathematisch-statistischer Modelle bei bewussten Auswahlen nicht möglich.

Willkürliche Stichproben: Elemente aus der Grundgesamtheit werden (etwa von einem Interviewer) mehr oder weniger willkürlich in die Stichprobe aufgenommen, es liegt ausschließlich im Ermessen des Interviewers oder auch der Untersuchungspersonen selbst.

5 Erstellen einer eigenen Statistik

Versuchen Sie, so gut wie Ihnen möglich, die folgenden fünf Kriterien zu erfüllen: „Objektivität“ (Unabhängigkeit vom Standpunkt des Statistikerstellers), „Reliabilität“ (Verlässlichkeit), „Validität“ (überkontextuelle Gültigkeit), „Signifikanz“ (Bedeutsamkeit) und „Relevanz“ (Wichtigkeit).

5.1 Aufgabe 1: Erstellen einer Körpergrössen-Handflächen Statistik

Messen Sie alle Körpergrössen und Handflächen in der Klasse aus.Wir möchten es ganz genau wissen. Präsentieren Sie Ihre Resultate grafisch sowie mündlich. Gibt es eine Korrelation zwischen den beiden erhobenen Datensätzen? Zusatz: Wie verhalten sich die Daten in Abhängigkeit mit dem Alter der Personen?

5.2 Aufgabe 2: Erstellen einer Raucher und Drogen Statistik

Erheben Sie Daten. Ziehen Sie alle möglichen Möglichkeiten in Betracht, denn wir wollen es ganz genau wissen. Wieviel wird an der Kantonsschule Luzern geraucht? Vieviele Drogen werden an der Kantonsschule Luzern konsumiert? Präsentieren Sie Ihre Resultate mündlich und grafisch und verweisen Sie auf alle Schwierigkeiten, die Sie bewältigen mussten.

6 Korrelation, lineare Regression, R-Software

6.1 Korrelation

6.1.1 Definition

Die Korrelation ist eine Beziehung zwischen zwei oder mehr statistischen Variablen. Wenn sie besteht, ist noch nicht gesagt, ob eine Größe die andere kausal beeinflusst, ob beide von einer dritten Größe kausal abhängen oder ob sich überhaupt ein Kausalzusammenhang folgern lässt.

6.1.2 Genauere Beschreibung

Es gibt positive und negative Korrelationen. Ein Beispiel für eine positive Korrelation (je mehr, desto mehr) ist: Je mehr Futter, desto dickere Kühe. Ein Beispiel für eine negative Korrelation (je mehr, desto weniger) ist: „Je mehr zurückgelegte Strecke mit dem Auto, desto weniger Treibstoff ist vorhanden.“ Häufig benutzt man zu Recht die Korrelation, um einen Hinweis darauf zu bekommen, ob zwei statistische Größen ursächlich miteinander zusammenhängen. Das funktioniert immer dann besonders gut, wenn beide Größen durch eine „Je...desto“ Beziehung miteinander zusammenhängen und eine der Größen nur von der anderen Größe abhängt. Beispielsweise kann man unter bestimmten Bedingungen nachweisen, dass Getreide umso besser gedeiht, je mehr man es bewässert. Hängt die Menge oder Qualität des Getreides jedoch zusätzlich zum Wasser noch von anderen Variablen ab (beispielsweise von der Temperatur, dem Nährstoffgehalt des Bodens, dem einfallenden Licht usw.), „verwäscht“ der kausale Zusammenhang in der Statistik immer mehr, falls nicht gleichzeitig diese Variablen auch untersucht werden. Die Korrelation beschreibt aber nicht unbedingt eine Ursache-Wirkungs-Beziehung in die eine oder andere Richtung. So darf man über die Tatsache, dass man Feuerwehren oft bei Bränden findet, nicht folgern, dass Feuerwehren die Ursachen für Brände seien. Die direkte Kausalität kann auch gänzlich fehlen. So kann es durchaus eine Korrelation zwischen dem Rückgang der Störche im Burgenland und einem Rückgang der Anzahl Neugeborener geben, diese Ereignisse haben aber nichts miteinander zu tun – weder bringen Störche Kinder noch umgekehrt. Das heißt, sie haben kausal allenfalls über eine dritte Größe etwas miteinander zu tun, etwa über die Verstädterung, die sowohl Nistplätze vernichtet als auch Kleinstfamilien fördert. Im Gegensatz zur Proportionalität ist die Korrelation nur ein statistischer Zusammenhang. Es kann nur eine ungefähre Zu- oder Abnahme prognostiziert werden. Eine 200-prozentige Steigerung der Futtermenge kann eine Gewichtszunahme der Kühe von 10 % oder auch von 20 % bewirken.

6.2 Lineare Regression: Methode der kleinsten Quadrate

Die Methode der kleinsten Quadrate (bezeichnender auch: der kleinsten Fehlerquadrate; englisch: Least Squares Method) ist das mathematische Standardverfahren zur Ausgleichungsrechnung. Es ist eine Wolke aus Datenpunkten gegeben, die physikalische Messwerte, wirtschaftliche Größen oder ähnliches repräsentieren können. In diese Punktwolke soll eine möglichst genau passende, parameterabhängige Modellkurve (z.B. eine Gerade) gelegt werden. Dazu bestimmt man die Parameter (im Falle der Gerade: Steigung a und der y-Achsenabstand b) dieser Kurve numerisch, indem *die Summe der quadratischen Abweichungen der Kurve von den beobachteten Punkten minimiert wird.*

6.2.1 Gerade durch drei Punkte?

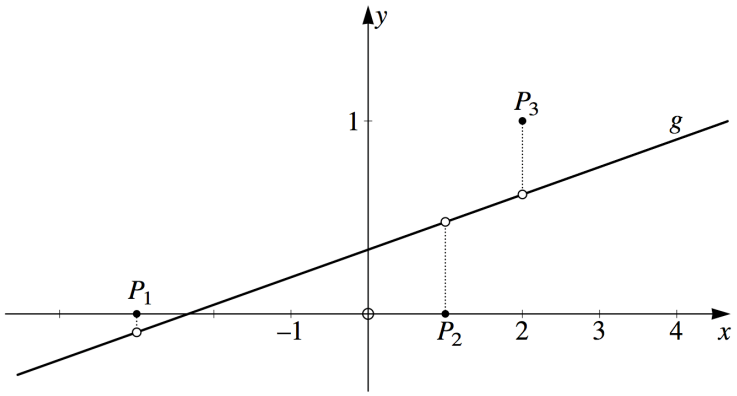


Abbildung 6.1: Gerade durch drei Punkte?

Der Ansatz $y = ax + b$ für die Geradengleichung enthält zwei freie Parameter a und b, die zu bestimmen sind. Wäre die Einsetzprobe für die Koordinaten der drei Punkte P_1 , P_2 und P_3 erfüllt, so würden die folgenden Beziehungen gelten:

$$\begin{aligned} 0 &= -3a + b \\ 0 &= a + b \\ 0 &= 2a + b \end{aligned}$$

oder mit den Abkürzungen

$$\vec{1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \vec{x} = \begin{pmatrix} -3 \\ 1 \\ 2 \end{pmatrix}, \qquad \vec{y} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

die Vektorgleichung $\vec{y} = a\vec{x} + b\vec{1}$. Widersprüche verhindern die Lösbarkeit dieser Gleichungen. Aber die Ordinatenwerte y_i der drei Punkte P_i dürfen verändert werden, um eine Lösung zu erzwingen. Es ist, genauer gesagt, ein Vektor $\vec{y}$ gesucht, der in der Ebene ϵ aller Linearkombinationen von $\vec{1}$ und $\vec{x}$ liegt und für welchen die Länge des Differenzvektors $\vec{r} = \vec{y} - \vec{y}\,'$ minimal wird da $\|\vec{r}\| \geqslant 0$ gilt, befindet sich das Minimum von $\|\vec{r}\|$ an derselben Stelle wie jenes von

$$\|\vec{r}\|^2 = \|\vec{r}\| \cdot \|\vec{r}\| = \sum_{i=1}^{3} r_i^2,$$

wobei angenommen wurde, dass r_i kartesische Koordinaten von $\vec{r}$ bezeichnen. Also ist die *Normalprojektion* $\vec{y}\,'$ von $\vec{y}$ auf ϵ der *beste Kompromiss im Sinne der Methode der kleinsten Quadrate*.

6.2.2 Die Lösung des Minimierungsproblems im Sinne der Methode der kleinsten Quadrate

In unserem Falle ist der lineare Ansatz $y_i = a + bx_i$. Indem man die x_i zur Datenmatrix A, die Parameter a und b zum Parametervektor $\vec{p} = (a, b)^T$ und die Beobachtungen y_i zum Vektor $\vec{y}$ zusammenfasst, kann man das lineare Gleichungssystem in Matrixform darstellen und der kleinste-Quadrate-Ansatz führt dann wieder wie oben auf ein lineares Ausgleichsproblem der Form

$$\min_{a,b} \left\| \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} - \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \right\|_2 = \min_{\vec{p}} \|\vec{y} - A\vec{p}\|_2.$$

Für die resultierende Ausgleichsgerade dieses einfachen (aber durchaus relevanten) Beispiels lassen sich die Lösungen für die Parameter direkt angeben als

$$b = \frac{(\sum_{i=1}^{3} x_i y_i) - 3\bar{x}\bar{y}}{(\sum_{i=1}^{3} x_i^2) - 3(\bar{x})^2} \qquad \text{und} \qquad a = \bar{y} - b\bar{x}$$

mit $\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i$ als arithmetisches Mittel der $\vec{x}$-Werte ($\bar{y}$ entsprechend).

Die Lösung für b kann mit Hilfe des Verschiebungssatzes auch als

$$b = \frac{\sum_{i=1}^{3}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{3}(x_i - \bar{x})^2}$$

angegeben werden.

Unser Minimierungsproblem hat immer eine eindeutige Lösung wenn die Matrix A vollen Rang hat. Die partiellen Ableitungen bezüglich der p_i und Nullsetzen derselben zum Bestim-

men des Minimums ergeben ein lineares System von Normalgleichungen (auch Normalengleichungen) ergeben die schönste und kürzeste Form:

$$A^T A\vec{p} = A^T\vec{y} \qquad \text{resp.} \qquad \vec{p} = (A^T A)^{-1}A^T\vec{y}.$$

6.3 R-Software

Das R-Package ist eine Software für: Datenkompilation, Statistische Datenanalyse und graphische Darstellung von Datensätzen und analytischen Resultaten. Alle standard Routinen sind implementiert (z.B. auch die lineare Regression im Sinne der Methode der kleinsten Quadrate).

6.3.1 Deskriptive Statistik:

- summary statistics: `summary`
- sample range: `range`
- sample mean: `mean`
- sample standard deviation: `sd`
- sample variance: `var`
- sample correlation matrix: `cor`
- sample quantiles: `quantile`

Examples:

```
y <- c(3,5,2,6,4,2,7,4,3,3,4)
summary(y); range(y); mean(y); sd(y); var(y);
quantile(y,seq(0,1,by=0.05))
```

6.3.2 Regression

- linear regression: `lm`
- nonlinear regression: `nls`
- generic functions on results: `summary, residuals, predict, coefficients`

Examples:

```
x <- c(0,1,2,3,4,5,6,7,8,9,10)
y <- c(3,5,2,6,4,2,7,4,3,3,4)
res.lm <- lm(y ~ x); summary(res.lm)
```

```
residuals(res.lm)
predict(res.lm,interval="confidence")
coefficients(res.lm)
summary(res.lm)$coefficients
```

6.3.3 Univariate Wahrscheinlichkeitsverteilungen:

- normal: `norm`
- log-normal: `lnorm`
- beta: `beta`
- gamma: `gamma`
- Student's t: `t`
- uniform: `unif`
- etc.

Example:

```
x <- rnorm(1000,0,1); hist(x,freq=F)
lines(seq(-3,3,by=0.1),dnorm(seq(-3,3,by=0.1),0,1))
```

6.3.4 Aufgabe in R:

Versuchen Sie die korrelierenden Daten der Körpergrösse-Handflächen Statistik zu auf ein lineares Modell zu fitten. Benutzen Sie dazu den vorbereiteten Skript:

```
#Eingabe
x <- c(-3,1,2)
y <- c(0,0,1)
#Verarbeitung
res.lm <- lm(y~x)
summary(res.lm)
#y=a+bx
a<-coef(res.lm)[1]
b<-coef(res.lm)[2]
gerade<- function (x) a+b*x
#Ausgabe
plot(x, y ,col=1 ,xlab ="x",ylab ="y")
x<-seq(min(x), max(x),length=100000)
curve(gerade(x), min(x), max(x), col=1, add = TRUE)
```

7 Wenn bei der Datenerhebung nur geschätzt werden kann

Dieses Kaptitel verlässt die Materie der deskriptiven Statistik. Ziel ist, unpräzises Wissen in Form von Wahrscheinlichkeiten zu erheben. Voraussetzungen für die Abhandlung sind Grundkenntnisse über kumulative Wahrscheinlichkeitsverteilungen $F(\theta) = P(\Theta < \theta)$ (CDFs cumulative distribution functions) und Wahrscheinlichkeitsdichten $f(\theta)$ mit $\int_{\theta_1}^{\theta_2} f(\theta)\, d\theta = P(\theta_1 < \Theta < \theta_2)$ (PDFs probability density funtions).

7.1 Erhebung von Quantilen

Indem man in der Gleichung der kumulativen Wahrscheinlichkeitsverteilung $F(\theta) = P(\Theta < \theta)$ den Wert P fixiert und den Experten nach einem geschätzten Parameterwert θ fragt, erhebt man sogenannte Quantile.

Der meist gefragte Parameterwert ist der Median, das ist wenn $P = 50\%$ beträgt. Weitere beliebte Quantile sind die Quartile, das ist wenn $P = 25\%$ bez. $P = 75\%$ beträgt.

Anbei die dazugehörigen Fragen:

Median Können Sie einen Wert θ_m bestimmen so dass Θ dieselbe Chance hat kleiner bzw. grösser als diesen Wert zu sein?

unteres Quartil Nehmen wir an Θ ist kleiner als der Median. Können Sie einen neuen Wert $\theta_l q$ angeben, für den die Chance für Θ gleich gross ist kleiner bzw. grösser als diesen Wert zu sein?

oberes Quartil Nehmen wir an Θ ist grösser als der Median. Können Sie einen neuen Wert $\theta_l q$ angeben, für den die Chance für Θ gleich gross ist kleiner bzw. grösser als diesen Wert zu sein?

7.2 Heuristiken

Heuristik (altgr. heurísko „ich finde“; heuriskein, „(auf-)finden“, „entdecken“) bezeichnet die Kunst, wahre Aussagen zu finden, im Unterschied zur Logik, die lehrt, wahre Aussagen zu begründen. Gerd Gigerenzer definiert wie folgt: Als Heuristik bezeichnet man eine Methode,

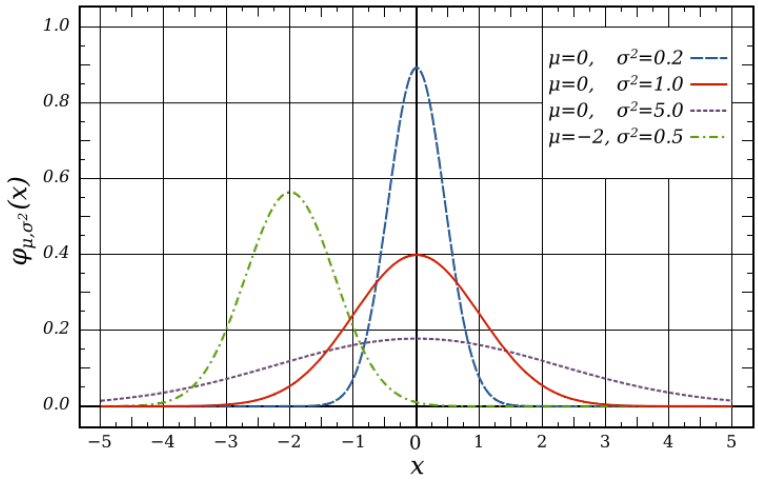


Abbildung 7.1: Dichtefunktion der Normalverteilung $N(\mu = 0, \sigma = 1)$

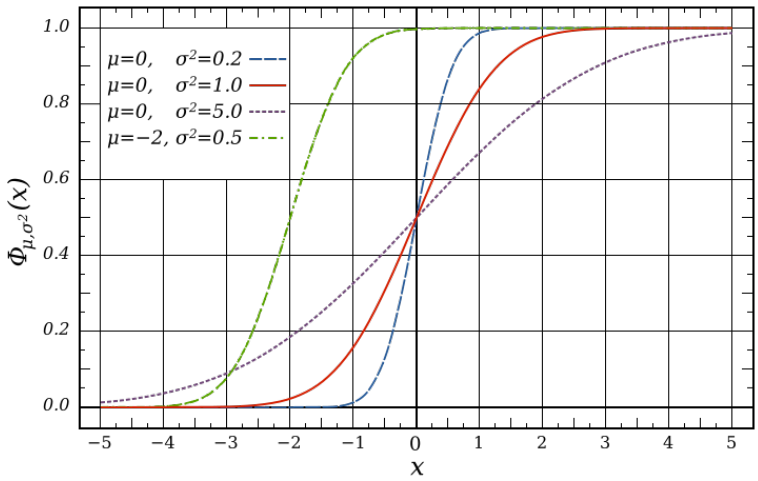


Abbildung 7.2: kumulative Wahrscheinlichkeitsverteilungen der Normalverteilung $N(\mu = 0, \sigma = 1)$

komplexe Probleme, die sich nicht vollständig lösen lassen, mit Hilfe einfacher Regeln und unter Zuhilfenahme nur weniger Informationen zu entwirren.

7.2.1 Die Ankerheuristik

Mit Anker- und Anpassungsheuristik bezeichnet man in der Sozialpsychologie eine unbewusste mentale Abkürzung, bei der sich das Urteil an einem beliebigen (willkürlichen) Anker orientiert. Die Folge ist eine systematische Verzerrung in Richtung des Ankers. Beispiel von Daniel Kahneman: Wenn ein Publikum zuerst gebeten wird, die letzten vier Ziffern der eigenen Sozialversicherungsnummer auswendigzulernen, und dann die Anzahl der ärzte in New York schätzen soll, dann beträgt die Korrelation beider Zahlen etwa 0.4 - weit mehr als dem Zufall entsprechen würde! An die erste Zahl nur zu denken, beeinflusst die zweite, obwohl es keine logische Verbindung zwischen beiden gibt.

Nicht nur Zahlen können als Anker dienen, sondern auch persönliche Erfahrungen und Beobachtungen.

7.2.2 Die Verfügbarkeitsheuristik

Die Verfügbarkeitsheuristik (engl. Availability heuristic) gehört in der Sozialpsychologie zu den sog. Urteilsheuristiken, welche gewissermaßen Faustregeln darstellen, um Sachverhalte auch dann beurteilen zu können, wenn kein Zugang zu präzisen Informationen besteht. Die Bezeichnung Verfügbarkeitsfehler (engl. Availability error) ist ebenfalls gebräuchlich für die dem Spielerfehlschluss verwandte Wahrnehmungsverzerrung. Sie beruht auf der Tendenz, bestimmte Ereignisse höher zu gewichten und eher in Erinnerung zu rufen als andere Ereignisse.

7.2.3 Die Repräsentativitätsheuristik

Die Repräsentativitätsheuristik gehört in der Sozialpsychologie zu den Urteilsheuristiken, welche gewissermaßen Faustregeln darstellen, um trotz großer Unsicherheit in Situationen zu schnellen und ökonomischen Urteilen zu kommen. Je ähnlicher eine Person einem typischen Vertreter einer bestimmten Gruppe ist, desto eher ordnet man die Person dieser Gruppe zu. In einer klassischen Untersuchung boten Daniel Kahneman und Amos Tversky (1983) ihren Versuchspersonen die schriftliche Beschreibung einer weiblichen Person namens Linda dar. Darin wurde sehr viel über Lindas Tätigkeit für Frauenrechte und Emanzipation berichtet. Danach wurden die Probanden gefragt, was denn nach dieser Beschreibung wahrscheinlicher sei, dass Linda ëine Bankangestellteöder ëine Bankangestellte und Feministinßei. Die Mehrzahl der Versuchspersonen schätzte die Wahrscheinlichkeit, dass Linda "Bankangestellte und Feministinßei, wesentlich höher ein. Diese Einschätzung ist jedoch irrig, denn die Wahrscheinlichkeit für das gleichzeitige Auftreten beider Ereignisse kann nicht größer sein, als die Wahrscheinlichkeit, dass eines der beiden Ereignisse alleine eintritt (Konjunktion, und-Verknüpfung).

(Selbst wenn alle Bankangestellten auch Feministinnen sind, wären die beiden Wahrscheinlichkeiten für (1) "Bankangestellteünd für (2) "Bankangestellte und Feministin"gleich groß.)

7.3 Aufgabe: Einschätzung der Anzahl Raucher pro Klasse an der Schule

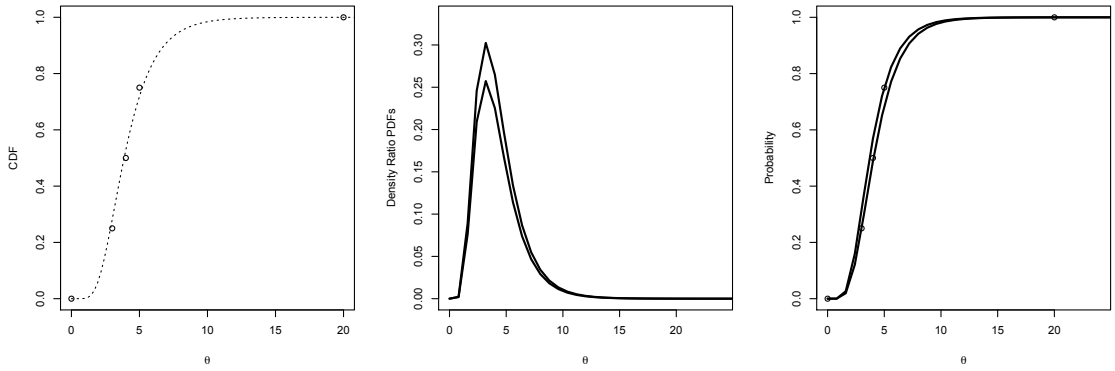


Abbildung 7.3: Gefittete Lognormalverteilung des Datensatzes: $x = (0, 3, 4, 5, 20)$, $y = (0, 0.25, 0.5, 0.75, 0.9999$. Der Plot entspricht dem untenstehenden R-Programm.

Schätzen Sie die Anzahl Raucher pro Klasse an der Schule. Erfragen Sie den Median, das untere und obere Quartil. Tun Sie das in Abhängigkeit des Jahrganges. Nehmen sie als Standardklassengrösse 20 Schüler. Vergleichen Sie die Ergebnisse mit der Raucherstatistik aus dem obigen Kapitel. Versuchen Sie die Unterschiede zu erklären.

Geben Sie die Daten in folgendes R-Programm ein, welches die Daten mit einer Lognormalverteilung fittet:

```
rm(list=ls())

#INPUT

#Quantiles
q <- c(0, 3, 4, 5, 20)
#Probabilities
p <- c(0, 0.25, 0.5, 0.75, 0.9999)
```

```
#PROCESSING

# LOG NORMAL Distribution Fit of Quantilefunction
qlnorm.nls <- nls(q ~ qlnorm(p, meanlog = A, sd = B),
start=list(A = 2, B = 1))

#Define Functions dnorm qnorm pnorm
fln<- function (x) dlnorm( x, meanlog=coef(qlnorm.nls)[1],
sdlog=coef(qlnorm.nls)[2])
Fln<- function (x) plnorm( x, meanlog=coef(qlnorm.nls)[1],
sdlog=coef(qlnorm.nls)[2])
qln<- function (x) qlnorm( x, meanlog=coef(qlnorm.nls)[1],
sdlog=coef(qlnorm.nls)[2])

#Calculation of Factor Kappa
x<-seq( q[1], q[5],length=100000)

# def FLower envelope
Fl<- function (x, k) (Fln(x)/(Fln(x)+k*(1-Fln(x))))
# def FUpper envelope
Fu<- function (x, k) (k*Fln(x)/(k*Fln(x)+(1-Fln(x))))

k<-1
k1.temp<-rep(1,length(q))
k2.temp<-rep(1,length(q))
#calculation of k
for(i in 1:length(q)){
if(p[i]<Fln(q[i])) {k1.temp[i]
<-(Fln(q[i])*(1-p[i]))/(p[i]*(1-Fln(q[i])))}
if (p[i]>Fln(q[i])){k2.temp[i]
<-(p[i]*(1-Fln(q[i])))/(Fln(q[i])*(1-p[i]))}
k<-max(k1.temp,k2.temp)}

#OUTPUT

par(mfrow=c(1,3))

#plot 1

#plot elicitated data
plot(q, p ,col=1 , xlab =(expression(theta)), ylab ="CDF")
```

```
#plot CDF of fitted Log-Normal-Distribution
curve(Fln(x), q[1]*1.2,q[5]*1.2, col=1, add = TRUE,lty="dotted")

#plot 2

#empty plot to set the second frame
plot(numeric(0),numeric(0),xlim=c(q[1]*1.2,q[5]*1.2)
ylim=c(0,k*max(fln(x))*1.1), xlab = (expression(theta)),
ylab ="Density Ratio PDFs")

#plot PDF of fitted Log-Normal Distribution
curve(fln(x), q[1]*4, q[5]*4, col=1, add = TRUE,lwd="2" )

#plot the unnormalized PDF of the fitted Log-Normal Distribution
curve(k*fln(x), q[1]*4, q[5]*4, col=1, add = TRUE,lwd="2")

#empty plot to set the fird frame
plot(numeric(0),numeric(0),xlim=c(q[1]*1.2,q[5]*1.2),ylim=c(0,1),
xlab = (expression(theta)), ylab ="Probability")

curve(Fl(x,k), q[1]*4, q[5]*4, col=1, add = TRUE,lwd="2")
curve(Fu(x,k), q[1]*4, q[5]*4, col=1, add = TRUE,lwd="2")
points(q, p ,col=1 , xlab =(expression(theta)), ylab ="")

#prints
#print fitting coefs of normal distribution
summary(qlnorm.nls)
#print final factor used
k
```