Brückenkurs Mathematik

T. von der Twer

Inhaltsverzeichnis

KAPITEL 1

Zur deskriptiven Statistik

1. Vorbemerkung

Eine Absicht dieses Kurses ist es, von der Mathematik her das *Verständnis* der Statistik zu unterstützen. Tatsächlich sind die Grundbegriffe der Statistik mathematisch, und ein Mangel an mathematischem Verständnis drückt sich unweigerlich in vielfältigen Mängeln aus bei dem Versuch, Statistik anzuwenden: Naheliegendes wird nicht angewandt; weicht eine reale Situation wesentlich von den Voraussetzungen ab, die für den Sinn einer Anwendung benötigt werden, so wird das missachtet; vollends gibt es Schwächen bei der Interpretation ausgerechneter Resultate: Sie gerät dürr oder gar unsinnig. Glücklicherweise ist die zur Grundlage benötigte Mathematik dem „gesunden Menschenverstand" zugänglich, nicht erst dem mathematischen Talent. Allerdings will die erstere durchaus verbreitete Gabe auch ein wenig mehr als alltäglich üblich strapaziert und geübt werden. Weiter fügt es sich günstig, dass dabei gleichzeitig auch die allgemeine Fähigkeit gestärkt wird, mit komplizierteren sowie abstrakteren Sachverhalten umzugehen - die abstrakten Grundbegriffe sowie quantitative Beschreibung als solche haben ihre Bedeutung weit über die Statistik hinaus. Insbesondere soll dieser Kurs verdeutlichen, dass Mathematik eine immense beschreibende Kraft besitzt (die ihre Erfolge bei Anwendungen auf nahezu allen Gebieten der Wissenschaft erklärt), und dass Mathematik nicht nur für speziell Begabte, die einen besondern Zugang zum Hermetischen haben, von Interesse ist.

Beginnen wir mit dem, was für einen Sozialwissenschaftler am nächsten liegt: Man möchte Eigenschaften einer sozialen Gruppe ausdrücken, beschreiben. Auf einer ersten begrifflichen Stufe liegen Merkmale der einzelnen Gruppenmitglieder, die dann (übertragen) eine Gruppeneigenschaft ergeben, Beispiele wären: hohes Sozialprestige, Bildungsniveau. Auf einer nächsten Stufe hat man Begriffe von Beziehungen (Relationen) zwischen Gruppenmitgliedern, z.B. „A hat höheren Rang als B" oder sympathischer „A kommuniziert intensiv mit B", usw. Solche Beziehungen haben im allgemeinen bereits interessante mathematische Strukturen, z.B. Symmetrie oder Ordnung. Hier wäre bereits ein tieferligendes *mathematisches* Resultat zu nennen: Mit Beziehungsbegriffen lässt sich *mehr ausdrücken* als mit einfachen (einstelligen) Eigenschaftsbegriffen, es handelt sich um eine echt höhere Komplexitätsstufe (die übrigens in der Mathematik viel früher realisiert und benutzt wurde als auf anderen Feldern, man denke an frühe Persönlichkeitspsychologie noch im 19. Jahrhundert, die nur von individuellen Eigenschaften als grundlegend und alle Beziehungen erzeugend sprach. Es ist aber im einzelnen ein interessantes Problem, welche Beziehungen sich aus den individuellen Eigenschaften ergeben, welche nicht.) Eine noch höhere Stufe wird bei entwickelter Wissenschaft stets wichtig: Beziehungen von Beziehungen: Wie sieht das Gefüge der Beziehungen in einer sozialen Gruppe aus, z.B.: Welche Eigenschaften einer Beziehung zieht welche

einer andern nach sich, usw. Oder: Wie sieht die gesamte Kommunikationsstruktur einer Gruppe aus? So gelangt man schließlich zur Betrachtung ganzer Strukturen (Mengen von Objekten mit ihren interessierenden Beziehungen) und zu Beziehungen zwischen Strukturen (etwa beim Verhältnis zweier sozialer Gruppen zueinander). Endlich werden umfassendere, abstraktere Strukturen gebildet, bei denen die Objekte bereits Strukturen sind. (Beispiel: Ein Netz von sozialen Institutionen.) Die Mathematik hat nicht nur zuerst einen solchen Begriffsaufbau benutzt, sie hat ihn auch völlig abstrakt dargestellt - für jeden derartigen Begriffsaufbau, von beliebigen Begriffen eines beliebigen Feldes - und wichtige Resultate dazu geliefert. Darüber hinaus werden mathematische Begriffe und Strukturen stets dann bedeutsam, wenn man eine komplexe Struktur genauer beschreiben möchte. Z.B. wird eine Analyse des Baus von Kommunikationsstrukturen ohne ausdrückliche Verwendung von Mathematik nicht gelingen. Stellen wir zur Anregung dies Problem: Man beschreibe Kommunikationsstrukturen geeignet so, dass darin Begriffe wie „zentralistischer Charakter" präzise und adäquat ausgedrückt werden können. Führen wir hier auch ein gelungenes Beispiel aus der Sozialwissenschaft selbst an: Arrow hat (um 1950) *mathematisch* bewiesen, dass es kein demokratisches Verfahren geben kann, aus den Prioritätenlisten der Mitglieder einer Gruppe eine Prioritätenliste der Gruppe zu machen, es sei denn, die Liste umfasst höchstens zwei Items (Punkte), etwa: „Kandidat A soll dies Amt eher übernehmen als Kandidat B". Ohne gehörige Mathematik wäre ein solches Resultat nicht zu erzielen gewesen. Eine wesentliche Leistung von Arrows bestand darin, vernünftige Minimalbedingungen für ein „demokratisches" Verfahren mathematisch zu formulieren, bei deren Verletzung von jedermann anzuerkennen ist, dass kein demokratisches Verfahren vorliegt.

Diese Bemerkungen sollten nur andeuten, dass man tendenziell für interessantere und tiefere Resultate mit der Mathematik zu tun bekommt, und ein Gegengewicht zur verbreiteten Auffassung bilden, Mathematik sei gerade für das Kompliziertere, Lebendigere nicht verwendbar. Diese Auffassung ist vielfach mit dem Aberwitz verbunden, bei der Rede von besonders Kompliziertem mit außerordentlich primitiven Strukturen auskommen zu können. Selbstverständlich müssen wir für diesen Kurs zu Einfacherem zurückschalten, und es lohnt sich durchaus, auf der ersten der genannten Stufen wieder zu beginnen, bei den (einstelligen) Eigenschaften von interessierenden Objekten, nicht nur wegen ihrer Einfachheit, sondern auch, weil es bereits hier eine notwendige und nicht selbstverständlich verfügbare Verfeinerung gibt: Von qualitativen Begriffen („weiblich-nicht weiblich", „klein-mittelgroß-groß") zu quantitativen Begriffen (Größen): Statt zu sagen, jemand sehe sehr viel fern, wäre es viel informativer, mitzuteilen, wie viele Stunden der betreffende Mensch im Mittel täglich fernsieht, noch informativer, wie sich das auf einzelne Sendungstypen aufgliedert. Variablen sind Merkmale, deren jeweilige Ausprägung durch eine Zahlangabe (oder auch eine geordnete Folge mehrerer Zahlangaben) zu beschreiben ist. Man beachte, dass qualitative Begriffe nicht etwa eine höhere Dignität besitzen, sondern einfach nur einen primitiveren Spezialfall darstellen: Für „weiblich-männlich" kommt man z.B. mit den Zahlen 1,0 aus, und alle Statistik, die man für Variablen betreibt, ist insbesondere auch für soch einfache Spezialfälle anwendbar und gültig. Was Variablen attraktiv macht für diesen Kurs, ist nicht nur ihre Allgegenwart in allen Bereichen und die innewohnende Kraft verfeinerter Beschreibung, sondern auch die Tatsache, dass sie selbst ein besonders einfaches Beispiel des allgemeinsten und nützlichsten Begriffs der Mathematik darstellen: Es handelt sich um den Begriff der Zuordnung (Abbildung, bei Zahlen auch

gern „Funktion" genannt). Bei der Variablen „Lebensalter (in einer Population von Menschen)" wird *jedem* Menschen dieser Population *sein* Lebensalter zugeordnet (in ganzen vollendeten Jahren oder feiner, wie man will). Ebenso für „Anzahl der Schuljahre", „Anzahl der (pro Jahr z.B.) gelesenen Bücher" usw. Welche fundamentale Rolle dieser Begriff der Zuordnung mittlerweile bei der Beschreibung von komplizierteren Sachverhalten spielt, ist kaum zu ermessen. Insbesondere können wir den gesamten Begriffsapparat der Statistik mit diesem Begriff aufbauen, beginnend mit dem grundlegenden Begriff der Statistik überhaupt: „Verteilung einer Variablen". Dabei ordnet man (im einfachsten Fall) jedem Zahlenwert die relative Häufigkeit zu, mit der er als Wert der Variablen auftritt. Es bereitet dem Anfänger gewöhnlich Schwierigkeiten, von einzelnen Zahlangaben zur Zuordnung aufzusteigen, aber dafür auch mit Erreichen dieser Stufe ein großer Schritt getan. Der Zugang ist eröffnet zur Beschreibung zeitlicher Entwicklungen, auch anschaulich durch Kurven, allgemeiner zur Beschreibung von Zusammenhängen zwischen Variablen (die man idealtypisch durch mathematische Funktionen beschreibt und deren Abweichung vom Idealtypus man wiederum mit Statistik in den Griff bekommt), in denen sich bereits sehr viele interessante Sachverhalte ausdrücken. Wir steigen damit wieder auf zu den Beziehungen von Beziehungen. Im einzelnen sollen die Beziehungen zwischen Stichproben und Gesamtpopulation und die „Korrelation von Variablen" genauer betrachtet und ausgeführt werden. Insgesamt sollte sich ein einigermaßen rundes und sicheres Bild von den Elementen quantitativer Beschreibung ergeben, die in der sozialwissenschaftlichen Literatur am häufigsten anzutreffen sind.

2. Der Begriff „Variable" (oder „Größe")

2.1. Merkmale und Merkmalsausprägungen. Merkmale sozialer Individuen sind z.B.: Alter, Familienstand, Schulabschluss, politische Einstellung, Bildungsstand, berufliche Qualifikation, wirtschaftlicher Status, Einkommen, Freizeitgestaltung usw. Alle diese Merkmale haben *verschiedene Ausprägungen*: Das Alter eines Individuums wird man in Jahren angeben (oder noch feiner durch gebrochene Jahreszahlen beschreiben - es macht Sinn, zu sagen, ein Individuum sei 3 Jahre älter als ein anderes. Beim Schulabschluss nur einen Typ angeben (Hauptschule, Realschule,...), dabei hat man noch eine vernünftige Ordnungsbeziehung: Ein Schulabschluss kann als „höher" gelten denn ein anderer; indessen ist es nicht eben sinnvoll, eine quantitative Differenz zwischen zwei Schulabschlüssen anzugeben. Bei der politischen Einstellung wird man völlig unterschiedliche Klassifikationssysteme haben, bei einzelnen darunter wird es noch eine solche Ordnungsbeziehung geben, bei anderen nicht: Man kann z.B. nicht die Parteien der BRD sinnvoll anordnen im Sinne einer von ihren Mitgliedern getragenen Einstellung. (Allenfalls kann man sie nach Anzahl ihrer Mitglieder anordnen usw.) Das Merkmal „Geschlecht" etwa hat nur ganze zwei Ausprägungen, keine Differenzbildung, keine Ordnung. Die angesprochenen drei Typen von Merkmalen nennt man gern „metrisch" oder „intervallskaliert" (sinnvolle quantitative Differenzbildung bei den Ausprägungen, feinere Zahlendarstellung der Ausprägungen, „ordinalskaliert" (sinnvolle Anordnung der Ausprägungen) oder eben nur „nominalskaliert" (nur Klassenbildung mit „Namen" für die Klassen). Halten wir eine einfache Binsenweisheit fest: Auch die so „qualitativ" (nach bloßem Vorurteil daher gegenüber Quantitativem „minderwertigen" wirkenden Merkmale, die typisch nominalskaliert sind, erlauben zahlenmäßige

Beschreibung (Codierung) ihrer Ausprägungen, lediglich ist diese Beschreibung *primitiver*: Für das Geschlecht genügt z.B., Eins für „weiblich", Null für „männlich" (man könnte auch die Zahlen 2,3 nehmen oder 2,6). Lediglich machen weitergehende Begriffsbildungen dann vielfach keinen Sinn, etwa: Wie weit liege ich über dem durchschnittlichen Geschlecht? Dennoch erweist sich der Mittelwert 1/2 (wir bleiben bei 0,1 und setzen voraus, dass es ziemlich gleich viele Frauen und Männer gibt, vorausgesetzt, man schliesst einmal höhere Altersklassen aus) als sinnvoll: Ziehen wir zufällig eine Stichprobe von 1000 Leuten, so werden wir dabei ziemlich genau je zur Hälfte Frauen und Männer haben.

Aber betrachten wir noch Merkmale wie „Bildungsstand" oder „berufliche Qualifikation": Diese sind mehrdimensional, d.h. eine einzelne Ausprägung besteht in einer geordneten Folge einzelner Ausprägungen, so dass sie sinnvoll gerade nicht mit einer Zahl zu beschreiben sind, sondern allenfalls mit einer geordneten Folge von Zahlen. Bildung oder berufliche Qualifikation hat jeweils verschiedene „Sparten", und es ist dann nützlicher, getrennt die Ausprägungen in diesen Sparten zu beschreiben. Sprachliche Fähigkeiten, Wahrnehmungsfähigkeiten, Wissen und abstraktes Denkvermögen, sie alle bezogen auf recht verschiedenartige Felder der „Bildung" machen einen „Bildungsstand" aus. Als ein solches mehrdimensionales „Profil" wird man sinnvoll eine komplexe Ausprägung des Merkmals „Bildungsstand" auffassen. Es sei hier nur darauf hingewiesen, dass statistische Methoden durchaus auch für mehrdimensionale Merkmale existieren und im Keim darauf beruhen, dass man getrennt für die einzelnen Dimensionen die eindimensionalen Methoden benutzen kann, darüber hinaus jedoch auch Zusammenhänge zwischen den Ausprägungen in den einzelnen Dimensionen durch Betrachtung mehrdimensionaler Verteilungen erreicht. Ein Beispiel wird die Regression und Korrelation sein, die wir besprechen werden.

Halten wir fest: Die Ausprägungen eindimensionaler Merkmale lassen sich stets durch Zahlen beschreiben. Ob dabei feinere Strukturen sinnvoll sind, hängt von der Willkürlichkeit einer solchen Zahlbeschreibung ab. Die Ausprägungen mehrdimensionaler Merkmale lassen sich dagegen stets mit Folgen von Zahlenwerten beschreiben. Wir werden nun über weiteste Strecken bei eindimensionalen Merkmalen bleiben.

Nachdem wir den Schritt vollzogen haben, die Ausprägungen von Merkmalen durch Zahlen zu beschreiben, liegt es nahe, von dem schwerfälligen Sprachgebrauch „Merkmal", „Merkmalsausprägungen" überzugehen zu „Variable" (oder aus „Größe"), „Werte der Variablen - oder der Größe". Man achte jedoch auf die Unterscheidung und merke sich simpel: Eine Variable ist ein Ding, das mehrere Werte haben kann. Natürlich interessiert uns gerade die Variation der Werte - hätten alle Individuen denselben Wert, so könnte man damit nichts unterscheiden, also nichts Interessantes beschreiben. Unbeschadet dessen ist es mathematisch zweckmäßig, auch als Grenzfälle Variablen mit konstantem Wert zu betrachten, diese sind einfach günstig bei manchen Rechnungen. Wir schließen diesen Fall daher ein und nicht aus. Von überragender Bedeutung für das Grundverständnis ist die nun folgende wesentlich genauere mathematische Erklärung des Begriffs der Variablen.

2.2. Der mathematische Begriff der Variablen. Man beachte, dass „Variable" noch für Buchstaben in Formeln wie $(x+y)^2 = x^2+2xy+y^2$ gebraucht wird, dass dies jedoch ein *anderer* Begriff ist! Es gibt noch weitere mathematische Verwendungsweisen des Wortes „Variable" - man beachte: für verschiedene Begriffe!

Zum Beispiel kennen Sie „unabhängige" und „abhängige" Variable. Damit kommen wir der Sache schon näher, wir werden nämlich sehen, dass „Variablen" der Statistik (im Grundbegriff!) sinngemäß nichts anderes als *Beispiele* für abhängige Variablen sind. Dies ist nun ein ziemlich altväterlicher Sprachgebrauch, den wir sogleich moderner erklären werden, mit dem Abbildungsbegriff. (Später lasse man sich nicht dadurch verwirren, dass im Kontext mit Regression wieder „Variablen" der Statistik als abhängige und unabhängige zu betrachten sind. Vielmehr erkenne man darin gerade die Tugend der Mathematik, dass in hundertfältiger Variation ihre Grundbegriffe auch auf höhreren Ebenen immer wieder anwendbar sind!) Knüpfen wir an den Gebrauch von Zahlen zur Beschreibung von Merkmalsausprägungen: Jemand ist 20 Jahre alt (in ganzen vollendeten Jahren), sieht täglich im Durchschnitt (hoffentlich nur) eine halbe Stunde fern. Zu einem andern Menschen der zu betrachtenden Gruppe mögen andere Werte gehören. Das ergäbe eine lange Liste. Wir wollen darüber reden, was hier und in unzähligen weiteren Beispielen geschieht, wollen außerdem praktischere Information als derartig lange Listen bereitstellen. Dazu muss man nur dies erfassen: Es wird eine Menge von Objekten (etwa Menschen einer bestimmten Altersgruppe in einem Gebiet zu einer Zeit) betrachtet, in der Statistik „Population" genannt. *Jedem* dieser Objekte wird *genau ein* Zahlenwert zugeordnet (die Ausprägung des interessierenden Merkmals). Eine solche Zuordnung nennt man eine *Variable*. Man erfasst sie auf abstrakterer Ebene völlig mit drei Informationen:

- Definitionsbereich = Menge der Objekte, denen etwas zugeordnet wird
- Wertebereich, eine Menge von Objekten, zu der alle zugeordneten Objekte gehören (bei Variablen stets die Menge der reellen Zahlen, die wir $\mathbb{R}$ nennen)
- Eine genaue Zuordnungsvorschrift: Sie besagt als allgemeine Regel, welches Objekt jedem einzelnen Mitglied des Definitionsbereiches zugeordnet wird.

Die in der Mathematik übliche symbolische Beschreibung einer solchen *Zuordnung* oder *Abbildung* sieht so aus:

$$\begin{array}{rccc} f: & A & \to & B \\ & a & \mapsto & f(a) \end{array}$$

Die erste Zeile ist zu lesen: „f geht von A nach B", A ist also Definitionsbereich und B Wertebereich. Die zweite Zeile liest man: „(Dem beliebigen Element) $a \in A$ wird das Objekt $f(a)$ (lies: „f von a") zugeordnet. Man beachte vor allem den Unterschied zwischen f - das ist die Abbildung selbst - und $f(a)$, das ist ein Element aus B.

Definition 1. *Eine Abbildung von einer Menge A in eine Menge B ist eine Zuordnung, die jedem Element von A genau ein Element von B zuordnet.*

Erstes Beispiel:

$$\begin{array}{rccc} f: & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & x^2 \end{array}$$

Dies ist die Quadratfunktion, deren Graph die bekannte Parabel in Normalgestalt ergibt. Man beachte, daß diese zwei Zeilen die gesamte Liste $f(1) = 1, f(2) = 4, f(-3) = 9, f(2/3) = 4/9$,... erfasst, die niemals zu einem Ende käme, in der auch Einträge für Zahlen wie π nicht mit exaktem Wert anzugeben wären. Besonders wichtig ist es, dass auch allgemeinere Aussagen in der Liste stecken wie

$f(x^2+y) = (x^2+y)^2 = x^4 + 2x^2y + y^2$. (Letztere Gleichung mit binomischer Formel.)

Zweites Beispiel:

$$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & X(\omega) \end{array}$$

Dies ist keine konkrete Abbildung, sondern nur das definitorische Schema (mit den üblichen Bezeichnungen) für eine Variable im Sinne der Statistik. Ein konkretes Beispiel könnte so aussehen:

$$\begin{array}{rccl} \mathit{Alter}: & \mathbb{W} & \to & \mathbb{R} \\ & S & \mapsto & \mathit{Alter} \text{ von } S \text{ in voll. Jahren} \end{array}$$

(Der Definitionsbereich $\mathbb{W}$ wäre noch zu präzisieren, etwa: „Menge aller eingeschriebenen Wuppertaler Studenten am 16.10.2000". Hier wäre also $X = \mathit{Alter}$ in voll. Jahren in der genannten Population von Studenten - so ausführlich muss das auch jeder Sozialwissenschaftler in Worten angeben! Man merke sich: Der Definitionsbereich einer Variablen in der Statistik heißt „Population". Dabei ist es wichtig, „Alter" nicht als eine Zahl zu verstehen, sondern als die gesamte Zuordnung. Das Alter eines bestimmten Studenten ist $\mathit{Alter}(S)$ - im altertümlichen Sprachgebrauch wäre genauer $A(S)$ die abhängige Variable, und S wäre die unabhängige. Aber wir wollen das genauer und einfacher so verstehen: S bezeichnet ein *beliebiges* Element von $\mathbb{W}$, $A(S)$ ist dann eine eindeitig gestimmte Zahl. A ist dagegen *keine Zahl* und auch nicht etwa die Menge aller Zahlen $A(S)$ für S aus $\mathbb{W}$ - diese nennt man korrekt die Menge aller Werte von A oder Bild von A, und die Identifikation von A damit führt zu völlig Unsinnigem (ungeachtet dessen steht das so in miserablen deutschsprachigen Statistikbüchern). Man sieht den wichtigen Unterschied zwischen A und $A(S)$ sehr schön daran, dass es sinnvoll ist, von einem Mittelwert und einer Streuung von *Alter* als einer Variablen zu sprechen. Eine konstante Zahl dagegen hat allenfalls sich selbst als Mittelwert und Streuung Null, das gibt also keine Information extra. In den Sozialwissenschaften würde man das letzte Beispiel etwa so in Worten benennen (und dasselbe meinen, nämlich die Zuordnung und keine einzelne Zahl!): „Alter in vollendeten Jahren bei den Wuppertaler Studenten, die am 16.10.2000 eingeschrieben waren". Genau genommen wird allerdings A erst zu einer Variablen im Sinne der Statistik, indem man so etwas wie Zählungen von Häufigkeiten beginnt oder abstrakter Wahrscheinlichkeiten dafür angeben kann, dass Werte in gewissen Bereichen auftreten. Das ist gemeint mit der „Verteilung von A", worauf wir noch ausführlich kommen werden. Jedenfalls wird man sich nunmehr nicht mehr zu sehr wundern, dass man die Variablen in der Statistik oft unter dem ausführlicheren Namen der „Zufallsvariablen" oder „zufälligen Größe" findet. Diesem Schema folgen alle bereits gegebenen und alle nachfolgenden Beispiele. Wir heben daher noch einmal hervor:

DEFINITION 2. *Eine Variable ist eine Abbildung von irgendeiner Menge Ω in die Menge $\mathbb{R}$ der reellen Zahlen. Symbolisch: $X : \Omega \to \mathbb{R}$. Mit $X(\omega)$ wird der Wert der Variablen bei einem beliebigen Populationsmitglied ω aus Ω bezeichnet. (Zusatz: Zu einer Variablen im eigentlichen Sinne der Wahrscheinlichkeitstheorie oder Statistik wird X erst dadurch, dass man weiter noch einen Wahrscheinlichkeitsbegriff auf Ω hat. Eine einfache Realisierung eines solchen Begriffes bei endlicher Population*

Ω besteht darin, dass man die relativen Häufigkeiten betrachtet, mit denen Werte von X aus beliebigen Bereichen vorkommen, z.B. auch die relative Häufigkeit, mit der ganz bestimmte Werte in der Population vorkommen.)

Ein wichtiger Punkt sei noch hervorgehoben, der mir selbst bei aufrichtig bemühten und keineswegs unfähigen fortgeschrittenen Studenten noch wiederholt begegnete: Es genügt einfach nicht, eine Variable mit der Zuordnungsvorschrift allein zu definieren, etwa als „Alter in vollendeten Jahren": Es macht keinen Sinn, etwa die Menge aller Objekte zu betrachten, denen man ein Alter zuordnen kann, alle Tiere und Untertassen und die Planeten wären dabei. Es macht dagegen wohl einen Sinn, die Variablen „Alter in der Population der Discobesucher" und „Alter in der Population der Opernbesucher", sagen wir des letzten Jahres in der BRD zu unterscheiden und miteinander zu vergleichen, z.B. wird man drastisch verschiedene Mittelwerte feststellen, und das sagt etwas aus über das kulturelle Klima. Man mache sich klar: Wir haben in solchen Fällen mit zwei verschiedenen Variablen zu tun, und dafür genügen die verschiedenen Populationen, wenn auch die Zuordnungsvorschrift dieselbe ist. (Wenn wir zu den Stichproben kommen, müssen wir sogar die (neue) Population aller Stichproben (aus einer vorgegebenen Population) festen Umfangs ins Auge fassen!)

Wir benötigen noch wenige weitere Vorbereitungen in mathematischer Grundbegrifflichkeit:

2.3. Buchstaben in Ausdrücken, Formeln, Aussagen.

- Konstanten:

Man spricht von einem bestimmten Objekt, der Zahl 2, der Zahl π oder auch von irgendeiner fest ins Auge gefassten Zahl x_0, von der Menge $\mathbb{R}$ oder einer bestimmten Menge Ω. Damit werden Aussagen wie $\pi \in \mathbb{R}$, $2 + \pi = 5$ gebildet, die ganz bestimmte Bedeutung haben und damit in ihrem Wahrheitswert bestimmt sind (wahr oder falsch).

- Freie Variablen:

Diese muss man benutzen, wenn man über *alle* Objekte eines Bereiches reden möchte und etwa sagen will, dass eine Aussage für *alle* Objekte daraus (oder auch für *irgendeines*) gilt. Insbesondere findet man sie in allgemeingültigen Formeln und in Definitionen, die $kraft\ Definition$ allgemeingültig sind. Beispiele:

$$(x \pm y)^2 = x^2 \pm 2xy + y^2 \text{ (binomische Formeln)}$$

Die freien Variablen sind hier x und y - für sie darf man irgendwelche Rechenausdrücke für Zahlen einsetzen, was den Fall der Konstanten als Spezialfall einschließt. Diese Formeln sind aus Grundaxiomen (oder Spielregeln) für Addition und Multiplikation herleitbar, ein mathematisches *Resultat*, das man (sehr leicht) beweisen kann. Es gibt durchaus schwierig zu beweisende mathematische Resultate, die tiefe Einsichten bringen. Insbesondere werden wir später sehen, dass der „Zentrale Grenzwertsatz" (an dessen Beweis im Rahmen dieses Kurses überhaupt nicht zu denken wäre) die Grundlage für den Löwenanteil der statistischen Praxis bildet.

Eine *definitorische* Gleichung wie $f(x) = x^2$ $(x \in \mathbb{R})$ meint völlig analog: Für *alle* reellen Zahlen x wird definiert, dass $f(x)$ das Quadrat von x sei. Wieder darf man für x jeden Rechenausdruck für eine Zahl einsetzen.

Bemerkung: Eigentlich handelt es sich beim angesprochenen (und sehr verbreiteten) Gebrauch von freien Variablen um eine abgekürzte Form von Allaussagen, z.B. ist mit der binomischen Formel genauer und vollständiger die folgende Aussage gemeint: *Für alle* x, y gilt: $(x+y)^2 = x^2 + 2xy + y^2$.

- Gebundene Variablen:

Freie Variablen, die in den Bereich eines Allquantors („für alle") oder Existenzquantors („es gibt") kommen, werden dadurch gebunden. Sie bleiben von anderweitigen Einsetzungen in die Aussage unberührt. Ein praktisch häufig auftretender Fall ist die Verwendung der Indizes bei großen Summenzeichen:

$$y = \sum_{i=1}^{n} x_i$$

bedeutet: n ist als Konstante oder äußerer Parameter (s.u.) zu verstehen. Der Wert der Summe ergibt sich dadurch, dass man *alle* Indexzahlen $i = 1, 2, ..., n$ nimmt (Konvention ist es hier, dass nur *ganze* Zahlen in Frage kommen), für sie jeweils den zugehörigen x-Wert bildet, also $x_1, ..., x_n$, und dann all diese Werte addiert. Es ist also definitionsgemäß:

$$\sum_{i=1}^{n} x_i = x_1 + ... + x_n$$

Beispiel: Ist $x_1 = 10, x_2 = 12, x_3 = 14$, das könnten etwa die Semesterzahlen von Studienabsolventen sein, so ist $\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 10 + 12 + 14 = 36$. Der arithmetische Mittelwert der drei angegebenen Zahlwerte ist dann $\frac{1}{3} \sum_{i=1}^{3} x_i = 12$.

- Unbestimmte (oder Unbekannte)

Bisher haben wir hauptsächlich Gleichungen vom Typus „allgemeingültige (oder definierende) Formel" betrachtet, mit den dort wesentlichen freien Variablen. Es treten jedoch auch völlig andersartige Gleichungen auf, zum Beispiel: Man weiß von einer gesuchten Zahl, dass sie eine gewisse Bedingung erfüllt, wie $2x - 3 = 4x + 2$. (Zugrundeliegen könnte hier die Frage nach dem Schnittpunkt zweier Geraden.) Diese Gleichung gilt nicht etwa allgemein, sondern nur für einen einzigen Wert von x. Sie ist eine *Bestimmungsgleichung* für die Unbekannte x. Man erhält $x = -5/2$.

- Äußere Parameter:

Sie werden in einer besonders wichtigen und typischen Situation verwandt: Man hat ein allgemeines System, das verschiedenste Konkretisierungen zulässt. Die Konkretisierungen werden durch Festsetzen bestimmter Zahlenwerte (oder sonstiger Objekte) vorgenommen. Man erhält nach voller Konkretisierung ein ganz bestimmtes System der ins Auge gefassten Art. Die Buchstaben der Beschreibung des allgemeinen Systems, für die dabei einzusetzen ist, nennt man äußere Parameter.

Beispiel:

$$f(x) = mx + b$$

So lautet die Gleichung für eine beliebige lineare Funktion $\mathbb{R} \to \mathbb{R}$. Dabei ist für x jede beliebige reelle Zahl einzusetzen. Nach unserem bisher eingeführten

Sprachgebrauch ist x eine freie Variable (genauer eine unabhängige Variable bei einer Zuordnung). Was ist mit m und b? Auch dafür kann man beliebige Zahlen einsetzen. Entscheidend ist nur: m und b sind *vorher* zu fixieren, gleich zu Beginn. Sobald das geschehen ist, redet man von *einer* linearen Funktion, deren Graph eine bestimmte Gerade ist. Aus der Schar aller Geraden sondert man eine bestimmte aus. Man denkt sich, dies sei getan, m und b die zugehörigen Werte, und rechnet nun allgemein mit den Buchstaben m und b weiter. m und b sind in diesem Gedankengang *äußere Parameter.* (Man sollte wissen, dass m die Steigung angibt, b den Schnittpunkt mit der y-Achse.) Heben wir den Nutzen hervor, den man davon hat, nicht etwa nur mit konkreten Einsetzungen wie $m = 2, b = 3$ zu arbeiten:

Wir stellen folgende Aufgabe: Man schneide *jede beliebige* Gerade in der xy-Ebene, die nicht parallel zur x-Achse liegt, mit der x-Achse. Die Aufgabenformulierung macht klar, dass es nicht genügt, einen Spezialfall zu behandeln. Daher ist es nötig, allgemein mit den äußeren Parametern m und b zu rechnen, deren Werte man sich *in beliebiger Weise fixiert* vorzustellen hat. Die Überlegungen und Rechnungen gehen nicht anders, als hätte man eine Konstante mit einem Buchstaben bezeichnet, wie π. Lösung der Aufgabe:

$mx + b = 0$. Nach Voraussetzung ist $m \neq 0$, also liegt der eindeutig bestimmte Schnittpunkt der zu m und b gehörigen Geraden bei $x = -b/m$. Besonders wichtig zu beachten ist hier: x ist Unbekannte, nach der man die Gleichung auflösen möchte. m und b dagegen sind äußere Parameter, nach denen aufzulösen kompletter Unfug wäre, einfach inhaltlich nicht sinnvoll, wenn auch durchaus formal korrekt möglich.

Bemerkung: Für den Fall $m = 0$ erhalten wir auch etwas Sinnvolles: Zwei Fälle treten auf: Ist $b = 0$, so ist *jeder* Punkt $(x, 0)$, $x \in \mathbb{R}$, ein Schnittpunkt, während wir für $b \neq 0$ keinen Schnittpunkt erhalten. Dies ist typisch für Probleme mit äußeren Parametern: Es sind mehrere Fälle getrennt zu überlegen, die verschiedenartig gelagert sind.

Noch einmal zum Verständnis: Wir haben die Schar aller Geraden (nicht parallel zur y-Achse), und für jede Wahl der Werte der äußeren Parameter erhalten wir eine konkrete Gerade und damit eine konkrete Schnittaufgabe. „Durchziehen" der äußeren Parameter beim Rechnen ermöglicht es, die Aufgabe *auf einen Schlag* für alle Fälle zu lösen. Die *allgemeine Lösung* lautet $x(m, b) = -b/m$ $(m \neq 0)$. Sie hängt von den Parameterwerten ab. Man kann nun die Antwort für *jeden* Spezialfall durch Einsetzen in die Endformel erhalten, z.B. für $m = 2, b = 3$: $x(2, 3) = -3/2$.

Wir werden einige für die Statistik besonders wichtige Typen theoretischer Verteilungen kennenlernen: Dabei wird jedes Exemplar des jeweiligen Typs (z.B. „Normalverteilung", auch „Binomialverteilung") durch Festlegen der Werte zweier Parameter bestimmt.

Es seien hier noch einmal die wesentlichen Typen von Gleichungen und Formeln zusammengestellt, die bereits bei der Einführung der Rollen von Buchstaben angeführt wurden:

- Definierende Gleichungen (Formeln)
- Allgemeingültige Gleichungen (Formeln)
- Bestimmungsgleichungen

Dazu kommt ein weiterer Typus:

- Formeln, die unter einer bestimmten Interpretation gelten

Beispiel: $a^2 + b^2 = c^2$ ist zweifellos nicht allgemeingültig, wohl aber immer richtig bei der Interpretation, dass a und b die Längen der Katheten, c die Länge

der Hypotenuse eines rechtwinkligen Dreiecks sind. Analog verhält es sich bei physikalischen oder sonstigen naturwissenschaftlichen Formeln. Wir werden solchen Formeln begegnen, deren Allgemeingültigkeit an einer ganz bestimmten statistischen Interpretation der Grundsymbole hängt.

3. Die elementaren Grundbegriffe der deskriptiven Statistik

Eine Vorbemerkung zum Zusatz „deskriptiv" (d.h. beschreibend): Wir wollen nur die wesentlichen statistischen Eigenschaften einer Variablen (auf ihrer „Gesamtpopulation" oder einfach „Population" Ω beschreiben, vorerst noch nicht anhand unvollständiger Information durch eine Stichprobe auf diese Eigenschaften schließen („schließende Statistik" oder „Inferenzstatistik") - vgl. dazu Kapitel 3.

3.1. Variablen mit wenigen Werten und ihre Verteilung, Mittelwert und Streuung. Den wichtigsten Begriff haben wir bereits eingeführt: Statistik handelt von Variablen, also Abbildungen $X : \Omega \to \mathbb{R}$, wobei Ω eine beliebige Menge von irgendwelchen Objekten ist. (Alle zugehörigen Überlegungen lassen sich gewöhnlich recht einfach auf vektorielle Variablen verallgemeinern, bei denen die Werte eben nicht einfach Zahlen, sondern Vektoren sind.)

Wir bemerkten, dass die volle Information über eine Variable X gewöhnlich in einer fürchterlichen Liste $(\omega \mid X(\omega))$ bestünde, bei unendlich großem Ω schließlich überhaupt nicht gegeben werden könnte. Das ist ganz anders bei mathematischen Abbildungen, für die man einen Rechenausdruck besitzt, wie $f(x) = x^2$. Darin steckt *alle* Information über die Quadratfunktion. Es stellt sich daher folgendes

Problem: Kann man im Falle einer statistischen Variablen zu einer ökonomisch eleganten mathematischen Beschreibung gelangen?

Offenbar geht das nicht, solange man die *volle* Information fordert, z.B. welche Körperlänge jeder einzelne Bundesbürger hat. Man muss daher Information wegwerfen, die vielleicht nicht gar so wichtig ist. Genau dies bewirkt auf geeignete Weise die *entscheidende statistische Abstraktion:*

Man fragt nur noch danach, *welche Werte* von einer Variablen X *wie häufig* vorkommen. Diese Information nennt man *die Verteilung* von X.

Wir nehmen *für diesen Abschnitt* zwei wesentliche Beschränkungen für unsere Betrachtungen vor (erstere wurde schon im Titel vermerkt):

- Wir reden nur von der Verteilung einer Variablen X in der *Gesamtpopulation* Ω, *nicht von Stichproben.*
- Wir setzen voraus, dass der Definitionsbereich Ω von X *endlich* ist (Spezialfall von „diskret verteilten Variablen", bei denen die Werte diskret auseinander liegen, kein Kontinuum bilden).

DEFINITION 3. *Sei $X : \Omega \to \mathbb{R}$ eine Variable mit endlichem Ω. Dann ist die Verteilung von X definitionsgemäß folgende Abbildung:*

$$\begin{array}{rccl} f_X : & \mathbb{R} & \to & \mathbb{R} \\ & a & \mapsto & \textit{relative Häufigkeit, mit der } a \textit{ als } X\textit{-Wert vorkommt.} \end{array}$$

Eine solche Verteilung kann man mittels eines Stabdiagramms veranschaulichen. Man beachte: Nunmehr steht die unabhängige Variable a der Funktion f_X für eine beliebige reelle Zahl, und $f_X(a)$ ist die relative Häufigkeit, mit welcher dieser Wert a als Wert der Variablen X auftritt. Kommt a gar nicht vor, so ist die absolute Häufigkeit Null, also auch die relative - das ist die absolute geteilt durch die Anzahl aller Populationsmitglieder. (Die prozentuale Häufigkeit ist die relative

mal hundert, oder auch: Die relative Häufigkeit ist die Häufigkeit pro 1 wie die prozentuale die pro hundert ist.)

Beispiel: In einer bestimmten Population von Studenten könnte man folgende Verteilung der Semesterzahlen gefunden haben (Tabelle zu f_X, wobei X die Variable „Semesterzahl in dieser Poplation" ist):

Semesterzahl	0	1	2	3	4	5	6	7
relative Häufigkeit	0.18	0.13	0.1	0.1	0.09	0.08	0.08	0.07
Semesterzahl	8	9	10	11				
relative Häufigkeit	0.07	0.05	0.04	0.01				

Hier ist ein Stabdiagramm dazu:

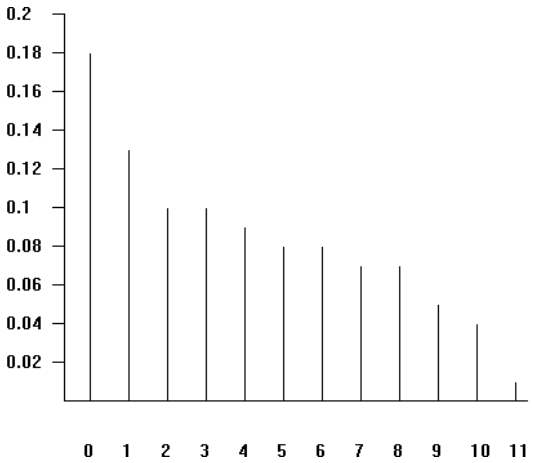


Abb. 1: Stabdiagramm der relativen Häufigkeiten, welche auf die einzelnen Variablenwerte entfallen - hier Semesterzahlen

Für manche Zwecke möchte man noch weiter reduzieren; tatsächlich genügen vielfach bereits zwei Zahlangaben, um eine Verteilung hinlänglich genau zu beschreiben, Mittelwert und Streuung. Der Mittelwert hat eine anschauliche Bedeutung im Rahmen eines Stabdiagramms: Man stelle sich die Stäbe vor mit einem Gewicht, das proportional zu ihrer Länge ist. Der Punkt auf der Größenachse (waagerechten Achse, also der Achse für die Werte der Variablen), an dem man unterstützen muss, um das Ganze im Gleichgewicht zu halten, das ist der Mittelwert. Rechnerisch bekommt man ihn so, dass man alle Einzelwerte mit ihrer relativen Häufigkeit multipliziert, das Ganze dann addiert. Dies kann man auch noch auf andere Weise verstehen: Man nehme die einzelnen X-Werte, zu den einzelnen Populationsmitgliedern, mit ihren Wiederholungen, bilde von diesen Werten das arithmetische Mittel, also die Summe von allen durch den Populationsumfang geteilt. Dann kommt dasselbe Resultat; denn in unserer Liste der relativen Häufigkeiten hat man einfach die wiederholten Werte zusammengefasst. Beispiel:

Mit Einzelwerten 0,0,0,1,1,2,2,3,3 hätte man arithmetisches Mittel $12/9 = 4/3$. Die Werte 0,1,2,3 hätten der Reihe nach relative Häufigkeiten: 1/3, 2/9, 2/9, 2/9, mit der Summe der Produkte „Wert mal relative Häufigkeit" erhielte man: $0 \cdot$

$1/3 + 1 \cdot 2/9 + 2 \cdot 2/9 + 3 \cdot 2/9 = 12/9 = 4/3$. Die Übereinstimmung beruht auf dem Distributivgesetz: $(3 + 3)/9 = (2 \cdot 3)/9 = 3 \cdot (2/9)$. Dies ist der Beitrag zum Mittelwert, der von den beiden Dreien kommt, zuerst in der Version, wie man das arithmetische Mittel der Einzelwerte berechnet, zuletzt in der Version: Wert mal relative Häufigkeit.

Wir fassen in einer definitorischen Formel zusammen:

DEFINITION 4 (Mittelwert einer Variablen). *Der Mittelwert einer Variablen X mit der Verteilung f_X lautet:*

$$\mu(X) = \sum_{X-Werte\ a} a \cdot f_X(a).$$ *(Auch einfach nur „μ", wenn die Variable klar ist.)*
Bemerkung: Sind mit x_i, $i = 1...n$, alle einzelnen Werte der Populationsmitglieder mit Wiederholungen aufgezählt, so gilt:
$$\mu(X) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Es sei bemerkt, dass die erste Formel auch dann verwendbar ist, wenn es nur endlich viele $X-$ Werte a gibt, für welche $f_X(a) \neq 0$. Insbesondere kann die Formel auch verallgemeinert angewandt werden, wenn es sich um Wahrscheinlichkeiten (statt relative Häufigkeiten) handelt. In beiden Hinsichten ist die zweite Formel weniger allgemein, sie setzt sowohl die Endlichkeit der Population als auch (beim Arbeiten mit Wahrscheinlichkeiten) die gleiche Wahrscheinlichkeit für jedes Populationsmitglied voraus. (Man vergleiche spätere Ausführungen, Stichworte: Zufällige Variablen, Verteilungen mit Wahrscheinlichkeitsdichten.)

In unserem tabellarischen Semesterzahl-Beispiel erhält man den Wert $\mu = 3.88$.

Nun zur zweiten charakteristischen Zahl für eine Verteilung: Die Streuung ist ein Maß für die Breite der Verteilung, die sich anschaulich auch als Breite des Stabdiagrammbildes darstellt. Aber man nimmt dazu nicht etwa die Breite des Intervalls, in dem die Werte liegen, sondern man gewichtet wieder: Wenige „Ausreißer" erhöhen die Streuung nicht nennenswert, wohl aber das Auftreten entlegener Werte mit nennenswerter relativer Häufigkeit. Intuitiv läge es nahe, den mittleren absoluten Abstand der Einzelwerte vom Mittelwert zu nehmen, doch wählt man in aller Regel ein etwas anderes Maß: Man bildet den Mittelwert der *quadrierten* Differenzen zum Mittelwert, anschließend zieht man die Wurzel, um ein Maß zu bekommen, das als Breite auf der Größenachse interpretierbar ist. Der Grund ist folgender: Wenn man gewisse Elemente der Wahrscheinlichkeitsrechnung und Statistik mathematisch tiefer durchdenkt, so stößt man bei der Beschreibung von mathematisch idealen Verteilungen wie Normalverteilungen und weiteren sehr häufig auf diese „Streuung" als Parameter. Sie ist also systematisch wichtiger. Im übrigen haben quadratische Differenzen noch den Vorteil der Differenzierbarkeit, was sie auch sonst für Fehlermessungen günstiger macht als absolute Differenzen.

Wiederum fassen wir in einer definitorischen Formel zusammen:

DEFINITION 5 (Varianz und Streuung einer Variablen). *Eine Variable X mit der Verteilung f_X hat folgende Streuung:*

$$\sigma(X) = \sqrt{\sum_{X-Werte\ a} (a - \mu(X))^2 \cdot f_X(a)}.$$ *(Oder einfach nur „σ".)*
Bemerkung: Das Ganze ohne die Wurzel ist zuweilen nützlich:
$\sigma^2(X)$ heißt Varianz von X (symbolisch auch: „Var(X)").
Bemerkung: Mit den Einzelwerten $x_1, ..., x_n$ sieht die Varianz so aus:
$\sigma^2(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu(X))^2$.

(Dieselben Bemerkungen zum Gültigkeitsbereich wie bei der vorigen Definition.)

In unserem Tabellen-Beispiel erhält man: $\sigma^2(X) = 9.87$ (gerundet), und die Streuung ist $\sigma(X) = 3.14$.

Wir wollen etwas genauer charakterisieren, was die Streuung besagt. Zunächst einmal kann man leicht feststellen, dass der Übergang von einer Variablen X zur Variablen $2X$ bewirkt, dass die Streuung sich verdoppelt: $\sigma(2X) = 2\sigma(X)$. Das folgt sofort aus der Rechenformel. Das Verteilungsbild zur Variablen $2X$ ist gegenüber dem zur Verteilung von X einfach mit Faktor 2 längs der Größenachse (waagerechten Achse) gestreckt. σ ist also wirklich ein Maß für die Breite. Andererseits ist σ nicht in allen Fällen unmittelbar geometrisch interpretierbar. Genau gibt σ eine Beschränkung für die relative Häufigkeit der Variablenwerte, die einen vorgebbaren Mindestabstand von μ haben. Zum Beispiel können außerhalb des Intervalls $\mu \pm 2\sigma$ *höchstens* 25% der Population liegen, mit mathematischer Sicherheit. Allerdings kann man die Faustregel angeben, dass im allgemeinen sogar weniger als 10% außerhalb dieses Intervalls liegen, mithin über 90% im Bereich $\mu \pm 2\sigma$ liegen. Bei Normalverteilungen sind es sogar ziemlich genau 95%.

Somit gibt die Streuung (insbesondere) an, in welchem Bereich um den Mittelwert der Löwenanteil der Population liegt, nämlich $\mu \pm 2\sigma$; Achtung: 2σ, nicht σ! Diese Aussage ist viel nützlicher als die Angabe eines Riesenintervalls, in dem mit Sicherheit alle Variablenwerte liegen. Schauen wir zur Konkretisierung noch einmal an, was in unserem Beispiel herauskommt: Bei der betrachteten Semesterzahlverteilung reicht das Intervall $\mu \pm 2\sigma$ von 0 bis 10, diese einschließend. Nur der Wert 11 liegt außerhalb, also nur 1% der Population in diesem Fall.

Man bedenke jedoch stets, dass μ und σ im allgemeinen nicht ausreichen, die gesamte Verteilung zu rekonstruieren - dies ist nur zuweilen der Fall, wenn man schon weiß, dass es sich um eine Verteilung eines gewissen Typs handelt, die tatsächlich vollständig mit diesen Parametern bestimmt ist. So verhält es sich z.B. bei Normalverteilungen.

Hat man die einzelnen Werte $x_1, ..., x_n$ einer Variablen X bei verschiedenen Populationsmitgliedern beobachtet, so definiert man auch:

DEFINITION 6 (arithmetisches Mittel der Werte $x_1, ..., x_n$).

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + ... + x_n}{n}$$

Dass diese Definition formal genau wie die von $\mu(X)$ lautet, verführt den Anfänger immer wieder dazu, beide miteinander zu verwechseln. Man beachte dazu

genau: Die Bezeichnung $\overline{x}$ steht für das arithmetische Mittel einer bloßen Stichprobe aus der Population. Erst dann, wenn die Stichprobe (was realistisch in den meisten Fällen nicht so ist) die gesamte Population umfasst, fallen beide zusammen, im Normalfalle ist eine Stichprobe viel kleiner als die Population, und dann liefert $\overline{x}$ einen sehr wichtigen Schätzwert für $\mu(X)$, auf den wir noch sehr genau eingehen werden. (Analog wird ein aus einer Stichprobe zu gewinnender Schätzwert s für die Streuung einzuführen sein, dazu später mehr.) Gerade dann, wenn man „nur" eine Stichprobe besitzt und die Qualität dieses Schätzwertes $\overline{x}$ für $\mu(X)$ diskutiert, wird diese Unterscheidung unerlässlich. Weiter bereitet es dem Anfänger dann Schwierigkeiten, abstrakt genug zu denken, dass $\mu(X)$, also in Worten das „Populationsmittel" *existiert*, auch wenn man es mangels der Kenntnis aller Einzelwerte *nicht konkret kennt oder ausrechnen kann* (!).

3.2. Variablen mit vielen Werten oder sogar einem Kontinuum von Werten: Histogramme und die Idee der Dichte und Verteilungsfunktion. Wir sind noch nicht fertig mit der einfachsten Beschreibung von Verteilungen; was sollte man mit einem Stabdiagramm beginnen, wenn eine Variable sehr viele Werte annimmt? Nehmen wir den Extremfall, dass man in einer riesigen Population jeden Wert nur einmal bekommt. Dann haben alle diese Werte dieselbe winzige relative Häufigkeit, und man erhält eine Art Rasenteppich als Stabdiagramm, nur stehen die Grashalme verschieden dicht verschiedenen Stellen. Dies wäre nicht nur sehr aufwendig zu zeichnen (per Computer ginge es natürlich wieder leicht), man hätte auch wenig davon. Aber das Problem führt zu einem tieferen Begriff, der für die gesamte theoretische Wahrscheinlichkeitstheorie und Statistik von größter Bedeutung ist, zum Begriff der Dichte. Die folgenden praktischen Ausführungen wollen auch dorthin geleiten, nicht nur weitere graphische Darstellungen von Verteilungen einführen.

Ein Beispiel: Wenn wir von einer großen Population von Jugendlichen die mittlere tägliche Fernsehdauer (unser X hier) erhoben haben, so werden wir etwa gruppieren: 0 bis unter 1 Stunde, 1 bis unter 2 Stunden, usw., bis 5 bis unter 6 Stunden. (Realistisch würde man weiter gehen müssen.) Dann werden wir die relativen Häufigkeiten zu diesen Klassen bilden und etwa zu folgender (unrealistischen!) Tabelle kommen:

Klasse	$0- < 1$	$1- < 2$	$2- < 3$	$3- < 4$	$4- < 5$	$5- < 6$
relative Häufigk.	0.15	0.4	0.2	0.14	0.1	0.01

Solche Daten kommen sehr oft vor. Stellen wir zunächst klar, dass man bei solcher Gruppierung eine Vergröberung der Verteilung von X vorgenommen hat: Z.B. weiß man in unserem Falle nicht, welcher Populationsanteil auf das Stundenintervall von 0 bis 1/2 entfällt - man wird vermuten, dass es sich um weniger als 0.075 handelt, aber die Zahl kennt man nicht. Das Vorgehen besteht nun einfach darin: Man betrachtet vereinfachend die Intervalle als völlig gleichmäßig besetzt und stellt die gruppierte Verteilung dann in einem Histogramm dar, auf folgende Weise:

Histogrammkonstruktion zu gruppierten Daten:Über jedem der Intervalle der Klasseneinteilung errichtet man einen Kasten, dessen Höhe der zugehörigen relativen Häufigkeit geteilt durch Kastenbreite entspricht und als Dichte zu interpretieren

ist. Bei dieser Konstruktion entsprechen die relativen Häufigkeiten der Klassen den-Flächeninhalten der Kästen. (Den Kastenhöhen entsprechen sie zugleich nur dann, wenn die Kästen alle dieselbe Breite haben.)

In unserem Beispiel erhält man das folgende Bild.

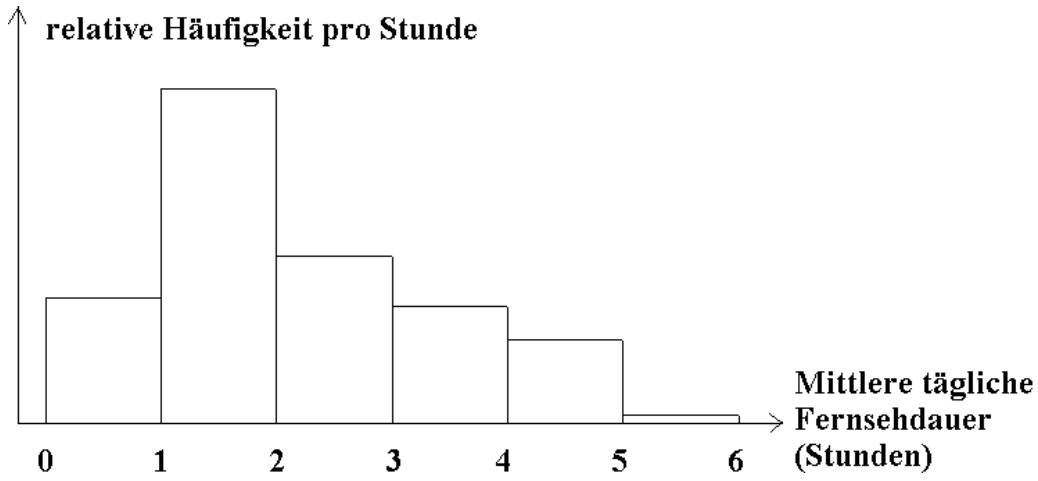


Abb. 2: Histogramm gruppierter Daten: Den relativen Häufigkeiten entsprechen die Flächeninhalte - im allgemeinen nicht die Höhen - das nur im Falle immer gleicher Klassenbreiten. Die Höhen repräsentieren Dichtewerte.

In *diesem* Falle könnte man wegen der gleichbleibenden Intervallbreite sogar auch die relativen Häufigkeiten der Tabelle ablesen. Anders sähe es aus, wenn man etwa in der Gruppierung die letzten beiden Intervalle zusammengefasst hätte zu: 4-6 Stunden: relative Häufigkeit 0.11. Ein Kasten über dem Intervall [4,6], dessen Höhe dem Wert 0.11 entspräche, würde den Anteil im Bild völlig verzerren, dem Intervall [4,6] mehr als das Doppelte von dem Gewicht geben, das es zuvor hatte. Die oben gegebene Vorschrift zur Konstruktion von Histogrammen korrigiert das automatisch: Division durch Klassenbreite ergibt eine Höhe, die dem Wert 0.055 entspricht, und dann erhält der Kasten über dem Intervall [4,6] denselben Flächeninhalt wie zuvor die beiden Kästen zu [4,5], [4,6] zusammen. Nun erkennt man auch, dass die Kastenhöhe allgemein abstrakter als Dichte zu interpretieren ist (hier: „relative Häufigkeit pro Jahr"). Der Wert 0.055 ist genau diese Dichte im Bereich [4,6], und er ist nicht als relative Häufigkeit deutbar.

Die Idee der Dichte führt auch theoretisch weiter: Wir können die Beschränkung auf endlich viele Werte und gar endliche Populationen fallenlassen und uns vorstellen, dass *jeder* Wert möglich ist; allerdings können wir nicht mehr von relativen Häufigkeiten reden, mit denen *einzelne* Werte vorkommen, sondern nur noch von relativen Häufigkeiten (oder besser noch: Wahrscheinlichkeiten) für *Intervalle (oder allgemeinere Bereiche)* von Werten. Für den einzelnen Wert haben wir stattdessen die Dichte. Weiter können wir verallgemeinern auf beliebige Dichtefunktionen - sie müssen nicht mehr stückweise konstant sein, sondern es können ihre Graphen beliebige Kurven sein. Nur sollte der Gesamtflächeninhalt unter der

Kurve existieren (als Integral) und den Wert 1 haben. Dann lassen sich als Flächeninhalte (Integrale) alle interessierenden Wahrscheinlichkeiten für beliebige Bereiche ausrechnen. Ebenso lassen sich μ und σ über Integrale berechnen. Tatsächlich spielt diese Verallgemeinerung eine immense theoretische Rolle: Wir werden sehen, dass man mathematisch gewisse „Idealverteilungen" (z.B. die Normalverteilungen) über Dichtefunktionen definieren und berechnen kann, die sich als außerordentlich wichtig auch für ganz praktische endliche Dinge erweisen, weil eben letztere in vielen Fällen systematisch voraussagbar sich einem solchen Idealtyp stark annähern. In diesem Zusammenhang wird eine weitere Darstellung wichtig, die *kumulierte.* Ihre besondere Tugend besteht darin, *universell* anwendbar zu sein, völlig gleichgültig, ob es sich um eine Variable mit diskreter oder kontinuierlicher Verteilung handelt, oder konkreter gesagt, ob die Verteilung durch ein Stabdiagramm oder eine Dichte gegeben ist. Dabei werden die Wahrscheinlichkeiten (oder relativen Häufigkeiten) aufsummiert, und zwar so: Man ordnet jeder reellen Zahl a die relative Häufigkeit (oder allgemeiner Wahrscheinlichkeit) für Werte $\underline{\leq a}$ zu. Das ist die entscheidende kleine Veränderung gegenüber der früher betrachteten Verteilung f_X. Bei ihr stand „=" statt „≤". Übrigens ist diese kumulierte Verteilung, die man „Verteilungsfunktion" nennt (das ist also ein terminus technicus!), auch manchmal in praktischer Hinsicht hilfreich: Histogramme eignen sich nicht bei Intervallen sehr verschiedener Breiten, da wegen der Flächendarstellung der Häufigkeiten Längen nicht verzerrt werden dürfen. Bei der Verteilungsfunktion stellen wieder die (an der vertikalen Achse abzulesenden) Funktionswerte die relativen Häufigkeiten dar, und die Intervallbreiten auf der Abszisse (horizontalen Achse) dürfen beliebig ungleichmäßig gedehnt oder gestaucht werden. (Denken Sie etwa an die Mitgliederzahlen in der Population der Vereine, da kommen sehr kleine und riesige vor. Das wird man nur mit der Verteilungsfunktion gut darstellen können.) Wir heben die Definition der Verteilungsfunktion zu einer Variable noch einmal in systematisch vollständiger Form hervor und geben anschließend wesentlich verschiedene Beispiele für Tabellen und Graphen von Verteilungsfunktionen. Außerdem gelangen wir bei der Erläuterung des zweiten der Beispiele zum überaus wichtigen Zusammenhang zwischen Dichte und Verteilungsfunktion.

DEFINITION 7. *Sei X eine Variable mit $\underline{\text{beliebiger}}$ Verteilung. Dann ist die Verteilungsfunktion von X folgende Abbildung:*

$$\begin{array}{rccl} F_X: & \mathbb{R} & \to & \mathbb{R} \\ & a & \mapsto & \textit{relative Häufigkeit der } X-\textit{Werte} \leq a. \end{array}$$

(Allgemeiner steht „Wahrscheinlichkeit" für „relative Häufigkeit".)

Man sollte sich diese Definition merken! Eingangs wurde *untechnisch* formuliert, die Verteilung einer Variablen gebe die Information darüber, welche Werte wie häufig vorkommen. Anschließend haben wir Stabdiagramme und Histogramme kennengelernt, welche je auf ihre Weise diese Information geben und jeweils nur für bestimmte Klassen von Verteilungen brauchbar sind. Nunmehr verfügen wir mit der Verteilungsfunktion über eine *technische* Fassung des Begriffs der Verteilung, welche *für alle Fälle brauchbar* und überdies *stets praktisch geeignet graphisch darstellbar ist* (wie die folgenden Abbildungen zeigen werden). Das sollte man zu schätzen wissen.

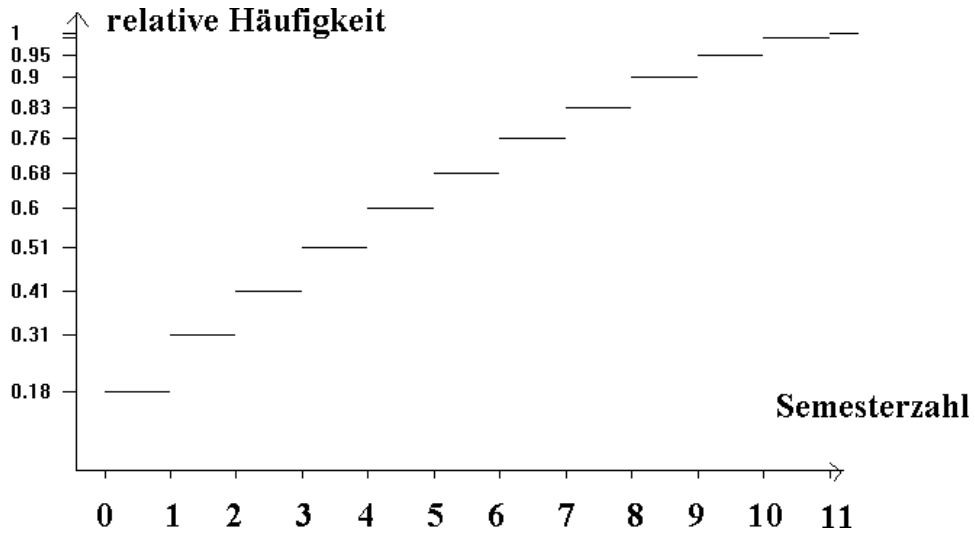


Abb. 3: Graphische Darstellung der Verteilungsfunktion zum Beispiel des ersten Stabdiagramms (Abb.1, S.11)

Zu beobachten sind die Treppen: Nur an den vorkommenden (hier ganzzahligen) Werten 0 bis 11 springt die Verteilungsfunktion, dazwischen kommt nichts hinzu, die Verteilungsfunktion bleibt also konstant. Vor dem ersten Wert (hier 0) hat sie konstant den Wert 0, nach dem letzten (11) konstant den Wert 1.

Das zweite Beispiel (Abb. 4 auf der nächsten Seite) zeigt die Verteilungsfunktion zum Histogramm von Abb. 2 (S. 13).

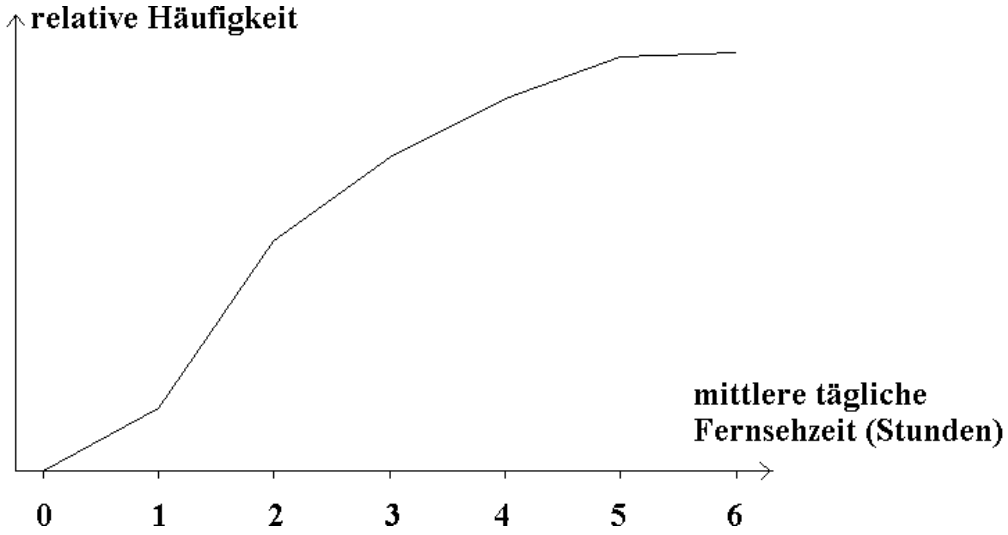


Abb. 4: Verteilungsfunktion zum früher gegebenen Histogramm (Abb. 2, S. 15)

Hier ist schon so etwas wie eine Kurve zu sehen, aber deutlich handelt es sich um einen stückweise geraden Polygonzug. Das liegt an dem linearen Anstieg innerhalb der Gruppierungsintervalle: Dort ist die Dichte konstant, also die Steigung der Verteilungsfunktion konstant. Dass die Dichte (das ist hier die stückweise konstante Funktion (graphisch: Treppe), welche die Histogrammkastenhöhen gibt) die Steigung der Verteilungsfunktion angibt, sieht man in unserem Beispiel folgendermaßen:

Nehmen wir das Intervall von 0 bis 1. Die Dichte ist konstant 0.15 (pro Stunde), da die Intervallbreite 1 beträgt. Wir haben $F_X(1) - F_X(0) = 0.15$, und $(F_X(1) - F_X(0))/1$ ist die *mittlere* Steigung von F_X auf [0,1] . Wenn man den $X - Wert$ von 0 bis 1 ansteigen lässt, kommt in gleichen Stücken stets der gleiche Flächeninhalt hinzu, da das Histogramm hier ein einziges Rechteck bildet. Folglich muss F_X auf [0,1] linear ansteigen, und 0.15, die mittlere Steigung von F_X auf [0,1], ist sogar die *konstante* Steigung der Verteilungsfunktion auf diesem Intervall. Das Histogramm gibt das Auf und Ab dieser stückweise konstanten Steigungen. Eine Dichte ist stets positiv (≥ 0), kann aber wachsen und fallen, eine Verteilungsfunktion hat stets Werte nur in [0,1] und ist stets monoton wachsend. Wir verdeutlichen noch einmal in allgemeinerer Form die Rolle der Klassenbreiten:

Sei eine Dichtefunktion auf dem Intervall [a,b] konstant, $a < b$, mit Wert c. (Man denke an [a,b] als ein Gruppierungsintervall (beliebiger Breite) bei einem Histogramm.

Dann ist c die Kastenhöhe:

$$c = \frac{relative\ Häufigkeit\ zu\ [a,b]}{b-a}$$

Aber der Zähler lässt sich auch ausdrücken als $F_X(b) - F_X(a)$, und es gilt:

$$Mittlere\ Steigung\ von\ F_X\ auf\ [a,b] = \frac{F_X(b) - F_X(a)}{b-a}$$

(Dies gilt allgemein für jede Funktion.)

Es kommt heraus, da diese mittlere Steigung zugleich die (konstante) Steigung von F_X auf [a,b] ist, an jeder Stelle des Intervalls:

SATZ 1. *Bei einer Verteilung mit (stückweise stetiger) Dichtefunktion f gilt für jede Stelle: Wert der Dichte = Steigung der Verteilungsfunktion, also:*

$$F'_X(a) = f(a)\ für\ jeden\ Wert\ a.$$

Umgekehrt ist die Verteilungsfunktion zur Dichte f daher diejenige Stammfunktion von f, deren Werte im Bereich [0,1] liegen.

Dies Resultat gilt völlig allgemein, auch für Dichtefunktionen, deren Graphen glatte Kurven sind. Idee: Man stelle sich Histogramme mit immer kleineren Intervallbreiten, immer feinerer Gruppierung vor, die sich der Kurve annähern. Aus der mittleren Steigung wird dann die Ableitung, die lokale Steigung in jedem einzelnen Punkt. Dies ist nichts anderes als der Inhalt des Hauptsatzes der Differential- und Integralrechnung, ausgesprochen speziell für positive Funktionen, die einen gesamten Flächeninhalt 1 mit der x-Achse bilden, d.h. Dichtefunktionen.

Nun wollen wir dies in Aktion sehen bei einer mathematisch idealen Verteilung, die zugleich die für praktische Zwecke wichtigste ist, der Standard-Normalverteilung

mit $\mu = 0$, $\sigma = 1$. Die folgende Graphik zeigt Dichte und Verteilungsfunktion zusammen.

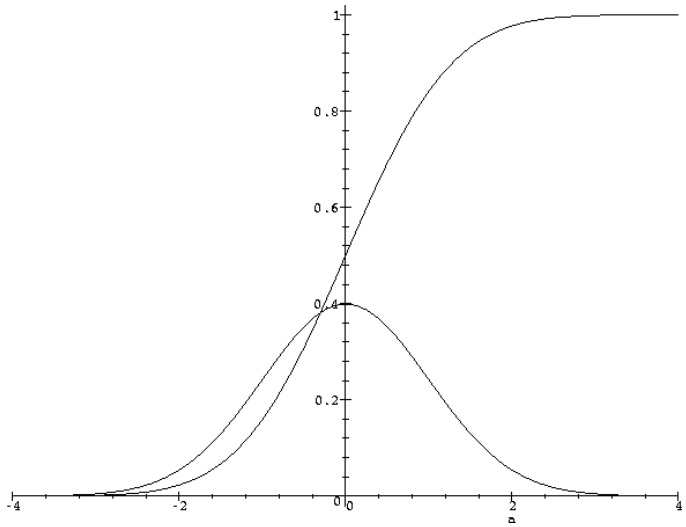


Abb. 5: Normalverteilung zu $\mu = 0, \sigma = 1$: Dichte (Glockenkurve) und Verteilungsfunktion (Sigmoid)

Abschließend schalten wir noch einmal zurück zu den Histogrammen und besprechen die Berechnung von μ und σ bei gruppierten Daten oder Histogrammen:

Bei gruppierten Daten kann man diese Verteilungsparameter nur näherungsweise angeben, und zwar zweckmäßig über die Idealisierung zu der Verteilung, die *genau* dem Histogramm entspricht, also mit stückweise konstanter Dichte. *Für diese Idealisierung* kann man leicht μ und σ *genau* angeben: Die Mittelwerte der Klassen entsprechen offenbar den Mittelpunkten der Klassenintervalle, und nun hat man (ganz gemäß Definition 3) einfach deren mit den relativen Häufigkeiten gewichtetes Mittel zu bilden, um μ zu erhalten. (Wenn man alle Einzelwerte, die in ein solches Intervall fallen, an der Intervallmitte ansiedeln würde, so erhielte man denselben Wert.) Wie steht es mit σ? Gewöhnlich macht man sich nicht einmal die Mühe, diesen Wert auch nur für die stückweise Gleichverteilung (also Histogrammverteilung) anzugeben, sondern benutzt einfach die Konstruktion, alle Einzelwerte in die jeweilige Klassenmitte zu verlegen. Dann ergibt die alte Formel: Varianz = mit den relativen Klassenhäufigkeiten gewichtetes Mittel der quadratischen Differenzen der Klassenmitten zu μ. Die Histogrammverteilung hat eine etwas größere Varianz als diesen Wert, da die quadratischen Entfernungen zur entlegeneren Intervallhälfte überwiegen. Aber für diesen Wert hätte man nicht nur komplizierter zu rechnen, er wäre auch in der Regel die schlechtere Näherung der Varianz der ursprünglichen Verteilung; denn meist hat man schiefe Glockenformen - so auch in unserem Beispiel, und dann gibt die Histogrammverteilung einen übertrieben großen Wert. Fassen wir zusammen:

Näherungsweise Berechnung von μ und σ bei gruppierten Verteilungsdaten:

$$\mu \approx \sum_{Klassen} Klassenmitte \cdot relative\ Häufigkeit\ zur\ Klasse$$

$$\sigma \approx \sqrt{\sum_{Klassen} (Klassenmitte - \mu)^2 \cdot relative\ Häufigkeit\ zur\ Klasse}$$

(Für μ ist hier natürlich der Näherungswert der 1. *Zeile einzusetzen.)*

3.3. Der Median als Alternative zum Mittelwert. In manchen Situationen erweist sich der Mittelwert als ungünstig für den Zweck der Beschreibung eines „Zentrums" einer Verteilung. Meist geht es dabei um das Problem weniger „Ausreißer", die durchaus (sogar einzeln!) in der Lage sind, das arithmetische Mittel gewaltig zu verschieben. Ein Beispiel: Nehmen wir eine Einkommensverteilung in einer Population von eintausend Menschen, deren Monatseinkommen sich mit sehr geringer Streuung um 4000 DM bewegen, sagen wir vereinfachend: Alle diese haben genau dies Einkommen. Nun fügen wir dieser Population einen einzigen Menschen hinzu mit Monatseinkommen 10 Millionen DM. Das arithmetische Mittel der Einkommensgröße wird dann von 4000 DM auf sage und schreibe knapp 14 000 DM angehoben! Natürlich sagt dieser Wert korrekt, was das arithmetische Mittel immer sagt, z.B., wie viel jeder Einzelne bekommen könnte, wenn gleichmäßig aufgeteilt würde. Aber man wird bemängeln, dass dieser Wert gerade im Niemandsland liegt, sogar eine überhaupt in der Bevölkerung nicht vertretene Größenordnung hat. Will man auf so etwas wie einen „typischen" Wert hinaus, dann ist bei sehr schiefen Verteilungen das arithmetische Mittel unbrauchbar. Gerade in solchen Situationen erweist sich der (oder besser „ein") Medianwert als günstig: Das ist ein (im allgemeinen nicht eindeutig bestimmter, aber doch hinreichend genau anzugebender) Wert auf der Größenskala mit der Eigenschaft: Die Hälfte der Population liegt unter diesem Wert, die andere darüber. In unserem Beispiel ist klar: Der Medianwert ist 4000 DM. Er gibt also das Typische für die Population - der „Ausreißer" beeinflusst diesen Wert überhaupt nicht. Generell gilt, dass einige Exoten kaum eine Veränderung im Median bewirken. Stets ist ein Median brauchbar als „typischer" Wert, *wenn* die Verteilung so etwas wie eine (eventuell schiefe) Glocke ist - und diese Form ist die bei weitem häufigste. Bei U-förmiger Verteilung ergibt natürlich auch ein Median nichts „Typisches".

Hier ein Beispiel zur Nichteindeutigkeit eines Medians: Liegen die Einzelwerte 0, 0, 1, 1, 1, 2, 2, 2, 3, 3 vor, so ist *jeder* Wert zwischen 1 und 2 als Median brauchbar. Man sollte dann auch nur sagen, der Median liege zwischen 1 und 2. Krampfhafte Versuche (sie existieren), mit irgendwelchen Regeln eine Eindeutigkeit zu erzwingen, sind völlig willkürlich und lohnen nicht. Bei einer ungeraden Zahl von Werten nimmt man den im Sinne der Anzahl „mittleren", also den sechsten bei 11 Werten, usw., aber auch hier könnte man sich getrost mit der Angabe eines Bereichs begnügen. In natürlicher Weise eindeutig definiert ist der Median bei einer Verteilung, die durch eine Dichte gegeben ist: Dann sind die relativen Häufigkeiten, die auf Bereiche von Größenwerten entfallen, durch die entsprechenden Flächeninhalte unter der Dichtekurve gegeben, und man findet genau eine Stelle, an welcher der gesamte Flächeninhalt unter der Dichtekurve halbiert wird. Dort hat die Verteilungsfunktion genau den Wert 0.5, der Median ist also definitionsgemäß exakt

diese Stelle. Für Verteilungen mit Dichtefunktion können wir daher den Median als (eindeutig bestimmte) Lösung der Gleichung $F_X(a) = 0.5$ definieren.

Es sei noch erwähnt, dass man bei „Mittelungen", die durch so etwas wie einen demokratischen Entscheidungsprozess entstehen, aus den erwähnten Gründen eher auf einen Median als auf den Mittelwert stoßen wird. Lassen wir zum Beispiel die Leute auf Zettel schreiben, wie viel Geld die Gesellschaft ausgeben sollte für einen bestimmten Zweck, dann ist die Sache nach unten begrenzt durch den Wert 0, nach oben aber offen. Exoten oder Witzbolde könnten erdenklich große Zahlen hinschreiben. Das arithmetische Mittel geriete in absurde Höhen. Ein Median wäre hier viel vernünftiger, auch „demokratischer", und tatsächlich wird so ein Wert als „gerechter" empfunden und setzt sich eher gesellschaftlich durch.

KAPITEL 2

Elementare Wahrscheinlichkeitsrechnung

1. Der Begriff der Wahrscheinlichkeit

1.1. Relative Häufigkeit und Wahrscheinlichkeit. Wir wollen den abstrakteren Begriff der Wahrscheinlichkeit vom vertrauteren der relativen Häufigkeit her entwickeln. Zunächst einmal brauchen wir einen generellen Rahmen, in dem eine Rede von „Wahrscheinlichkeit" erst sinnvoll werden kann: Stets muss ein (im Prinzip beliebig) wiederholbares Zufallsexperiment vorliegen. Im interessanten Fall sind bei der Durchführung eines solchen Experiments verschiedene möglich, wir begnügen uns zunächst mit endlich vielen Ausgängen. Zwei klassische Beispiele, an denen man die Sache gut verstehen kann:

Würfeln mit einem gewöhnlichen Würfel: Die möglichen Ausgänge sind die Augenzahlen 1,2,3,4,5,6.

Zufälliges Ziehen (also nach „blindem Mischen") einer Kugel aus einer Urne: Die möglichen Ausgänge sind die einzelnen Kugeln (die wir etwa numeriert haben mögen).

Beobachten Sie: Das zweite Beispiel enthält das erste strukturell als Spezialfall! (Man nehme eine Urne mit 6 Kugeln, und bei Wiederholen des Experiments ist die gezogene Kugel stets zurückzulegen und neu zu mischen.)

Bleiben wir beim zweiten Beispiel, das schon eine gewisse Allgemeinheit besitzt. Betrachten wir nun ein Merkmal in der Menge der Kugeln, allgemeiner gesprochen: bei den möglichen Ausgängen, zum Beispiel seien von den insgesamt n Kugeln k rot, der Rest weiß. Das Merkmal „rot" ist also in der Urne mit einer relativen Häufigkeit von k/n vertreten. Wenn man nun fragt, wie wahrscheinlich es nun sei, beim zufä lligen Ziehen einer Kugel eine rote zu ziehen, so wird fast jeder antworten, diese Wahrscheinlichkeit sei gerade diese relative Häufigkeit, k/n. Diese Antwort ist auch völlig korrekt. Aber nicht jedem ist dabei klar, was diese Aussage bedeutet. Was ist überhaupt gemeint mit: „Das Ereignis ... (das bei Durchführung eines bestimmten Experiments eintreten kann) hat die Wahrscheinlichkeit ... (eine Zahl aus [0,1])"? Jedenfalls meint diese Aussage nicht unmittelbar eine relative Häufigkeit, diese Übereinstimmung ist allenfalls ein *Resultat* mathematischer Überlegung. Sondern sie zielt auf *künftig (bei langer Wiederholung des Experiments) zu Erwartendes ab.* In unserm Beispiel besagt sie: Bei oftmaliger Wiederholung des Ziehens wird die relative Häufigkeit *der Fälle, in denen eine rote Kugel gezogen wurde (*unter allen Ziehungen), *nahe bei k/n liegen.* Dies ist eine Voraussage, und wir können empirisch prüfen und sogar mathematisch berechnen, wie gut sie ist. Wir halten dies als naive, informelle (aber überaus nützliche) Definition des Wahrscheinlichkeitsbegriffs fest:

Gegeben sei ein Zufallsexperiment, dazu ein Ereignis A, das bei diesem Experiment eintreten kann (oder auch vielleicht nicht). Dann gilt folgende Grundtatsache: Bei oftmaliger Durchführung des Experiments setzt sich eine relative Häufigkeit durch, mit der A beobachtet wird. Diese ideale relative Häufigkeit, um welche die beobachteten relativen Eintrittshäufgkeiten mit tendenziell immer kleineren Abständen pendeln, heißt Wahrscheinlichkeit von A, symbolisch: P(A).

Diese Definition ist auch in solchen Fällen brauchbar, in denen nicht bereits eine relative Häufigkeit zugrundeliegt, wie in unserem Urnenbeispiel.

Die beschriebene Grundtatsache wollen wir einmal am Werke sehen in einem Beispiel (vgl. Abb. 6, nächste Seite): Hier wurde (Computer-simuliert) 1000 mal gewürfelt und für jede Durchführungszahl i ($1 \leq i \leq 1000$ - waagerechte Achse) aufgetragen (senkrechte Achse), mit welcher relativen Häufigkeit eine gerade Augenzahl beobachtet wurde bis einschließlich zu diesem Versuch. Man sieht, dass die Wahrscheinlichkeit 1/2 so gut wie nie genau erreicht wird, die Abweichungen auch stellenweise wieder größer werden können, aber doch auf lange Sicht stets kleiner (tatsächlich sogar beliebig klein) werden. Dagegen sind bei wenigen Versuchen durchaus große Abweichungen zu möglich (und auch zu erwarten). Insbesondere kann die beobachtete relative Häufigkeit nach einem einzigen Versuch nur den Wert 0 oder aber 1 haben.

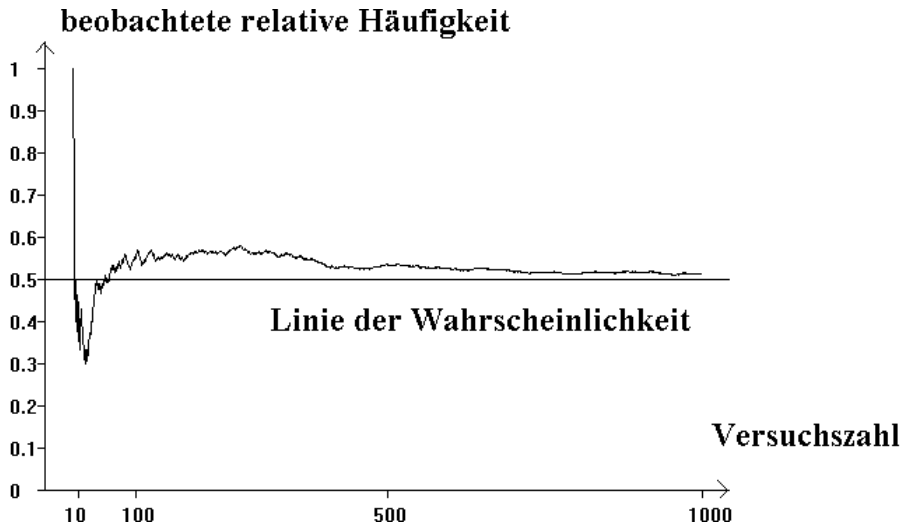


Abb. 6: Näherung empirischer relativer Häufigkeiten an eine ideale Wahrscheinlichkeit (hier mit Wert 1/2)

Die nächste Graphik zeigt noch etwas mehr: Dort wurden nicht nur mehrere Versuchsreihen aufgenommen, die immer das gleiche Bild zeigen - nur starten einige bei dem ganz falschen Wert 0, einige bei 1. Außerdem wurden Kurven mit eingezeichnet, die angeben, wie nahe bei der Wahrscheinlichkeit eine beobachtete relative Häufigkeit bei der jeweiligen Zahl der Versuche mit einer Wahrscheinlichkeit von 0.95 (oder 95%) liegen sollte. (Beobachten Sie, dass in einigen wenigen Fällen diese

Kurven überschritten wurden.) Zur Ermittlung dieser Kurven vgl. den späteren Abschnitt zur Anwendung der Normalverteilungen.

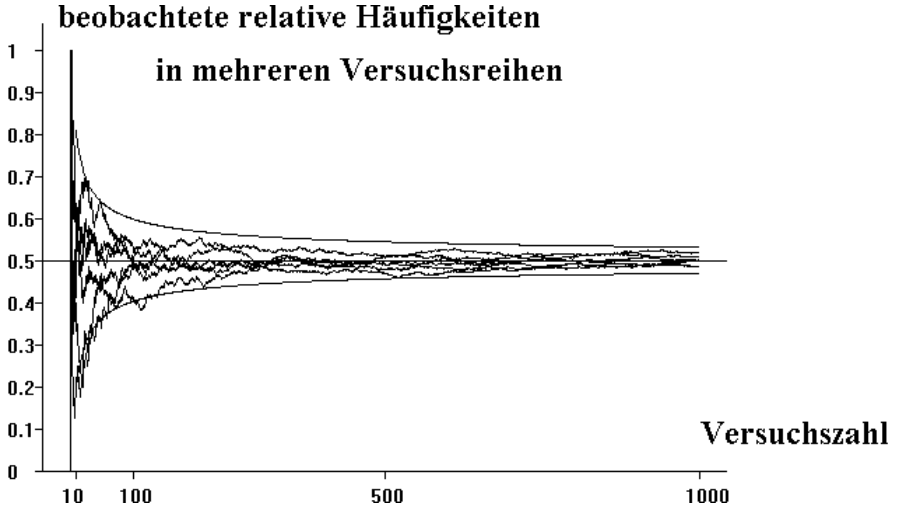


Abb. 7: Wahrscheinlichkeit und beobachtete relative Häufigkeiten, mit Begrenzungskurven, die mit 95% Wahrscheinlichkeit eingehalten werden.

Was wir hier beobachtet haben, ist die Wirksamkeit des „Gesetzes der großen Zahl", grob formuliert: Die empirischen relativen Häufigkeiten nähern sich (immer besser und sicherer) den Wahrscheinlichkeiten.

1.2. Der abstrakte Wahrscheinlichkeitsbegriff. Mathematisch geht man nunmehr so vor: Die Wahrscheinlichkeiten sollen so etwas wie „ideale relative Häufigkeiten" sein, daher muss man von ihnen verlangen, dass sie sich rechnerisch auch wie relative Häufigkeiten benehmen. Unmittelbar ergibt die in der folgenden Definition niedergelegten axiomatischen Anforderungen an den Wahrscheinlichkeitsbegriff, wenn man eine kleine geistige Vorbereitung getroffen hat: *Jedem Ereignis* soll *seine* Wahrscheinlichkeit (eine Zahl aus [0,1]) zugeordnet werden. Aber wie kann man *alle denkbaren Ereignisse* fassen? Es genügt, dies für ein beliebiges fest gegebenes Zufallsexperiment zu schaffen. Dabei hat man eine bestimmte Menge Ω möglicher Ausgänge. Nunmehr kann man jedes Ereignis so formulieren:

„Der zufällig herauskommende Ausgang ω hat die Eigenschaft E". Dies lässt sich umformulieren zu:

„Der zufällig herauskommende Ausgang ω ist Element der Menge $\{\omega \in \Omega \mid \omega$ hat die Eigenschaft E$\}$."

Einziger Bedeutungsträger in diesem Satz ist die Menge $\{\omega \in \Omega \mid \omega$ hat E$\}$. Somit kann jedes Ereignis als eine Teilmenge von Ω codiert werden, und mit der Menge aller teilmengen von Ω, $\mathcal{P}(\Omega)$, haben wir jedenfalls alle denkbaren Ereignisse erfasst.

Man prüfe das Verständnis dieser Bemerkungen, indem man folgende Ereignisse beim Würfeln (mit nur einem gewöhnlichen Würfel) verbal formuliert: $\{2, 4, 6\}, \{1\}$ und das Ereignis „Es kommt eine Zahl unter Drei heraus" als Menge umschreibt.

DEFINITION 8 (Begriff der Wahrscheinlichkeitsfunktion). *Sei Ω endlich. (Man denke an Ω als Menge aller möglichen Ausgänge eines Zufallsexperiments). Dann heißt eine Funktion P Wahrscheinlichkeitsfunktion über Ω, wenn sie folgende Eigenschaften besitzt:*

(1.1)
$$\boxed{\begin{array}{l} P : \mathcal{P}(\Omega) \to [0,1] \textit{(Lies P(A): „Wahrscheinlichkeit von A".)} \\ P(\Omega) = 1 \\ P(A \cup B) = P(A) + P(B), \textit{ wenn } A \cap B = \varnothing, \textit{ für alle } A, B \in \mathcal{P}(\Omega) \end{array}}$$

In unserem beschränkten Fall von endlichem Ω sind alle Teilmengen von Ω Ereignisse. Der Begriff ist jedoch sogar auf überabzählbare Ω verallgemeinerbar, allerdings kann P dann *nicht mehr auf der gesamten Potenzmenge von Ω definiert werden.* Insbesondere benötigt man die Verallgemeinerung der Summenformel auf abzählbar unendlich viele Mengen, die vereinigt werden, zu: $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$, wenn die Mengen A_i paarweise leeren Durchschnitt haben. (Auf der rechten Seite steht eine unendliche Reihe.)

Unmittelbar gewinnt man aus diesen Axiomen die Folgerungen:

(1.2)
$$\boxed{\begin{array}{ll} (i) & P(\overline{A}) = 1 - P(A) \\ (ii) & P(A \cap B) = P(A) + P(B) - P(A \cup B) \textit{ (stets!)} \\ (iii) & P(A) = \sum_{\omega \in \Omega} P(\{\omega\}) \textit{ (Für endliche, aber auch abzählbare } \Omega.) \\ (iv) & \begin{array}{l} \textit{Spezialfall von } (iii) : \textit{Sind alle Ausgänge gleich wahrscheinlich, so gilt:} \\ P(A) = \textit{relative Häufigkeit von } A \textit{ in } \Omega = \frac{\textit{Anzahl der „günstigen" Fälle}}{\textit{Anzahl der möglichen Fälle}} \end{array} \end{array}}$$

Insbesondere *(iii)* wirft ein Licht darauf, was am Wahrscheinlichkeitsbegriff unbefriedigend geklungen haben sollte: Er sagt nämlich nicht, wie man Wahrscheinlichkeiten ausrechnen kann, sondern nur, wie Wahrscheinlichkeiten miteinander zusammenhängen. Damit sind wir bei der geistigen Hauptfigur der Mathematik: Über die Zusammenhänge der Dinge miteinander das Wesentliche über diese Dinge herauszufinden. Aus den Zusammenhängen, die in den Axiomen stehen, folgert man mit (iii), dass man zum Ausrechnen von Wahrscheinlichkeiten lediglich die Wahrscheinlichkeiten der einzelnen Ausgänge selbst, genauer der Elementarereignisse $\{\omega\}$, zu kennen braucht. Nun sollte man verstehen, dass der *allgemeine* Wahrscheinlichkeitsbegriff nicht mehr sagen kann, da er z.B. für einen symmetrischen Würfel, aber auch noch für den schiefsten Würfel gelten sollte. Allenfalls bei zusätzlicher Information über den Einzelfall kann man erwarten, die Wahrscheinlichkeiten der Elementarereignisse spezifizieren zu können. Ein besonders einfacher Fall: Man hat endlich viele, sagen wir Anzahl n, die alle gleich wahrscheinlich sind. (Symmetrischer Würfel, „zufälliges Ziehen" usw.) Dann ist klar, dass alle die Wahrscheinlichkeit $1/n$ haben, und man gelangt zu (iv). Für kompliziertere Fälle liefert das Gesetz der großen Zahl einen empirischen Zugang: Man kann die relative Häufigkeit beobachten, mit der ein Ereignis eintritt, und damit die Wahrscheinlichkeit recht sicher und genau annähern. Die zwei grundlegenden theoretischen Methoden zur Gewinnung von Wahrscheinlichkeiten werden an den elementarsten und praktisch wichtigsten Beispielen in den Abschnitten 2.1 bis 2.4 dieses Kapitels gezeigt.

1.3. Von den Variablen zu den Zufallsvariablen (oder „zufälligen Größen"). Die meisten Wahrscheinlichkeiten, nach denen zu fragen interessant ist, beziehen sich auf Variaben, deren Werte man bei Zufallsexperimenten beobachten kann: Wie wahrscheinlich ist es, unter 1000 zufällig gezogenen Leuten wenigstens 300 mit einem gewissen Merkmal zu finden? Andererseits kann man auch Ereignisse, die nur eintreten oder nicht eintreten können, mit Werten 1,0 einer Größe beschreiben, so dass dieser Gesichtspunkt allgemeiner ist.

Wir stellen die begriffliche Verbindung her zwischen den eingangs betrachteten Variablen mit ihren Verteilungen (dort ging es um relative Häufigkeiten von Werten) zu den Zufallsvariablen, deren Werte man jeweils bei einem Zufallsexperiment beobachten kann, und ihren Verteilungen (dabei geht es um Wahrscheinlichkeiten von Werten). Dazu folgende

DEFINITION 9.

Eine Zufallsvariable X ist eine Abbildung $\Omega \to \mathbb{R}$, wobei über Ω eine Wahrscheinlichkeitsfunktion P gegeben ist.
Wenn Ω höchstens abzählbar ist, so ist die Verteilung von X ist folgende Abbildung:

$$\begin{array}{rccl} f_X: & \mathbb{R} & \to & \mathbb{R} \\ & a & \mapsto & P(X=a) := P(\{\omega \in \Omega \mid X(\omega) = a\}) \end{array}$$

Im ganz allgemeinen Fall hat man die stets brauchbare Verteilungsfunktion von X:

$$\begin{array}{rccl} F_X: & \mathbb{R} & \to & \mathbb{R} \\ & a & \mapsto & P(X \leq a) := P(\{\omega \in \Omega \mid X(\omega) \leq a\}) \end{array}$$

Man beachte: In den bisherigen Definitionen der Verteilung (nichtkumulativ bzw. kumulativ) ist jeweils einfach „relative Häufigkeit" durch „Wahrscheinlichkeit" zu ersetzen. Es ist sehr nützlich, sich zu merken, was Ausdrücke wie $P(X \leq a)$, $P(X > a)$, $P(X = a)$ bedeuten.

Zur Einübung in die Notation ein einfaches Beispiel: $X = Augensumme\ beim\ Würfeln\ mit\ zwei\ Würfeln.$ Wir haben hier:

$\Omega = \{(a, b) \in \mathbb{N} \times \mathbb{N} \mid 1 \leq a, b \leq 6\}$, also Menge der Paare natürlicher Zahlen bis 6.

$P(X = 2) = 1/36$, $P(X = 3) = 1/18$, $P(X = 4) = 1/12$, $P(X = 5) = 1/9$, $P(X = 6) = 5/36$, $P(X = 7) = 1/6$, $P(X = 8) = 5/36$, $P(X = 9) = 1/9$, $P(X = 10) = 1/12$, $P(X = 11) = 1/18$, $P(X = 12) = 1/36$.

2. Drei wichtige Verteilungstypen

Im ersten Kapitel haben wir den Begriff der Verteilung eingeführt, der sich vollkommen von relativen Häufigkeiten zu Wahrscheinlichkeiten verallgemeinert. Insbesondere verstehen wir nunmehr stets allgemein unter der Verteilungsfunktion F_X einer Zufallsvariablen X die Funktion von $\mathbb{R}$ nach $[0, 1]$ mit der definierenden Gleichung $F_X(a) = P(X \leq a)$, in Worten: $F_X(a)$ ist die Wahrscheinlichkeit dafür, dass ein $X-$ Wert $\leq a$ herauskommt bei Durchführung des zu X gehörigen Zufallsexperiments. Nun fragt sich, wie man an diese Wahrscheinlichkeiten herankommt. Hat man für das Zufallsexperiment ein genaues (oder näherungsweise zutreffendes mathematisches Modell, so kann man die zugehörigen allgemeinen mathematischen Resultate nutzen, in denen die Wahrscheinlichkeiten gerade zugänglich gemacht

werden (über Formeln, eventuell noch Tabellen, heutzutage vielfach in Computerprogrammen direkt nutzbar). (Andere Möglichkeiten bestehen im zusätzlichen oder auch alleinigen Einsatz empirischer Stichproben.) Wir stellen nunmehr die drei elementarsten mathematischen Modelle vor, die besonders häufig und auch vielfältig zu nutzen sind. Davon sind zwei Verteilungstypen diskret, in diesen Fällen mit nur endlich vielen möglichen Werten der Zufallsvariablen, welche so verteilt sind, dazu gesellt sich der wichtigste stetige (kontinuierliche) Verteilungstyp: Normalverteilung. So verteilte Variablen können **alle** reellen Zahlen als Werte haben. Das kommt so *genau* empirisch also nicht vor, dennoch ist dieser Typ der gerade für *praktische* Anwendungen wichtigste überhaupt, weil viele empirische Variablen diesem mathematischen Idealtyp *sehr nahe* kommen.

2.1. Binomialverteilungen. Eine Verteilungsform ist in der Regel auf einen bestimmten Situationstyp zugeschnitten, und das System (die Situation) ist zu konkretisieren durch einen oder mehrere Parameter. So auch in unserem Beispiel (und bei allen weiteren interessanten Verteilungstypen). Stellen wir folgendes Problem: Bei einem (beliebigen) Zufallsexperiment trete ein gewisses Ereignis mit Wahrscheinlichkeit p ein. Man führt das Experiment n mal durch. Mit welcher Wahrscheinlichkeit tritt dabei das Ereignis genau k mal ein? Diese Wahrscheinlichkeit hängt sicher von p, n, k ab. Aber es ist nützlich, p und n als äußere Parameter zu betrachten, k dagegen als unabhängige Variable; denn k ist ein beliebiger ganzzahliger Wert von 0 bis n, er variiert noch, wenn n und p bereits festgelegt sind. Zum Beispiel: Welche Wahrscheinlichkeit hat man für k Sechsen bei 100 Würfen? Hier sind p und n fixiert, aber k möchte man durchaus noch variieren. Die vollständige Antwort kann mit einer Formel gegeben werden. Aber solch eine Formel kann man immer nachschlagen - wichtiger ist das Verständnis der *Situationen*, in denen die Formel gültig ist. (Weiter unterscheide man dann noch, wann sie praktisch oder nur sehr mühsam bis unmöglich zu verwenden ist.)

DEFINITION 10.

Eine Variable X heißt p-Bernoulli-verteilt, wenn sie nur die Werte 1,0 annimmt, und zwar den Wert 1 mit Wahrscheinlichkeit p.
Eine Variable heißt $(n,p)-$binomialverteilt, wenn sie Summe von n unabhängigen p-Bernoulli-Variablen ist.
Wesentlich anschaulichere Beschreibung der binomialverteilten Variablen: Die (n,p)-binomialverteilten Variablen sind gerade die mit folgender Struktur: X = Trefferanzahl bei n unabhängigen Versuchen, wobei die Trefferwahrscheinlichkeit in jedem Versuch p beträgt.

Hier ist die Formel für die (nichtkumulierte) Verteilung:

SATZ 2. *Sei X (n,p)-binomialverteilt. Dann gilt:*

$$f_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ \text{für } k = 0, 1, ...n \tag{2.1}$$

Dabei ist das Symbol $\binom{n}{k}$ zu lesen : „n über k”,

$$\text{Berechnung} : \binom{n}{k} = \frac{n!}{k!(n-k)!}, \ \text{wobei } n! = 1 \cdot ... \cdot n \ (n \geq 1), \ 0! = 1,$$

lies „n Fakultät".

Beispiel: Mit welcher Wahrscheinlichkeit hat man bei 12 Würfen mit einem Würfel genau drei Sechsen? Die Antwort ist: $\binom{12}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^9 = 0.1974$.

Der Beweis der Formel ist nicht besonders schwierig. Man benötigt drei Überlegungen:

Bei unabhängigen Ereignissen ergibt sich die Wahrscheinlichkeit der „und"-Verbindung durch Multiplikation der Einzelwahrscheinlichkeiten. Beispiel: Die Wahrscheinlichkeit dafür, zwei mal hintereinander eine Sechs zu würfeln, ist $\frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$. Man beachte, wie sich die möglichen Fälle multiplizieren zu $6 \cdot 6 = 36$ und wie diese wieder gleich wahrscheinlich sind. Das ist das Wesentliche bei der Unabhängigkeit: Tritt das eine Ereignis ein, so ändert sich damit nicht die Wahrscheinlichkeit für das andere. Wenn man von einem Menschen das Geschlecht kennt, so ändern sich damit die Wahrscheinlichkeiten für gewisse Merkmale (Körperlängen, bevorzugter Typ von Literatur, auch von Sportarten), für andere nicht (IQ z.B.). Folglich ist $p^k(1-p)^{n-k}$ die Wahrscheinlichkeit dafür, die k „Treffer" und $n-k$ „Nieten" in einer ganz bestimmten Reihenfolge (gleichgültig welcher) zu erhalten.

Zweite Überlegung: Wie viele Anordnungen von k „Treffern" auf n Plätzen gibt es? (Gerade so oft hat man den Wert $p^k(1-p)^{n-k}$ zu nehmen, da sich diese Ereignisse, dass sich die Treffer in ganz bestimmter Anordnung ereignen, gegenseitig ausschließen und daher für die „oder"-Verbindung die Wahrscheinlichkeiten zu addieren sind (Axiom!). Wählen wir die k Plätze, so haben wir n Möglichkeiten für den ersten, dann unabhängig $n-1$ für den zweiten, zusammen also $n(n-1)$ Möglichkeiten, weiter geht es mit $n-2$ für den dritten usw., bis $n-k+1$ für den k-ten. Das macht $n(n-1)\cdot\ldots\cdot(n-k+1)$. Das ist jedoch noch nicht die gesuchte Zahl; wir müssen die Reihenfolge noch loswerden - wir wollten k Plätze, keinen ersten, zweiten, ..., k-ten, plastischer: k gleichberechtigte Vorsitzende aus n Leuten, nicht einen ersten, zweiten,..., k-ten Vorsitzenden.

Dritte Überlegung: Es bleibt noch zu teilen durch die Anzahl der Reihenfolgen, in die man k Objekte bringen kann. Das aber ist $k! = k(k-1)\cdot\ldots\cdot 1$, also das Produkt aller Zahlen von 1 bis k ($k \geq 1$). Nehmen wir nämlich an, wir wüssten schon die Anzahl der Möglichkeiten, m Dinge anzuordnen. Überlegen wir dann, wie viele es für $m+1$ sind: Fixieren wir eines der Objekte - wir können es als erstes ... bis letztes nehmen. Das sind $m+1$ Möglichkeiten. Unabhängig können wir die restlichen m Objekte *untereinander* auf so viele Weisen anordnen, wie wir schon wissen. Somit haben wir folgendes Resultat:

Anzahl der Möglichkeiten, $m+1$ Dinge anzuordnen = $(m+1)\cdot$ Anzahl der Möglichkeiten, m Dinge anzuordnen. Für $m = 1$ kommt eine Anordnung, für $m = 2$ kommen also $2 \cdot 1 = 2$ mögliche Anordnungen, für $m = 3$ werden es also $3 \cdot 2$. Allgemein für k also $k!$ mögliche Anordnungen. (Diese Schlussweise, von einer ganzen Anfangszahl mit einem Schema von jeder beliebigen zur nachfolgenden überzugehen und damit das Resultat für alle ganzen Zahlen von der Anfangszahl ab zu haben, nennt man *vollständige Induktion.)* Auch der Grenzfall $0! = 1$ stimmt, es gibt bei rechter Deutung *eine* Möglichkeit, die leere Menge von Objekten anzuordnen: Keiner ist der Erste/Letzte, also im Nichtstun besteht das Anordnen. Zusammen haben wir folgenden

SATZ 3. $\binom{n}{k}$ = *Anzahl der Möglichkeiten, k Dinge aus n Dingen auszuwählen. Dabei ist* $0 \leq k \leq n$, *k und n sind als ganze Zahlen vorauszusetzen.*

Dieser Satz gilt auch für $k = 0$. Denn es gibt genau eine Möglichkeit, 0 Dinge aus k Dingen auszuwählen - man lässt alle ungewählt. Man könnte auch sagen, dass man alle n als Liegengelassene auswählt. Und offensichtlich gibt es genau eine Möglichkeit, n Dinge aus n Dingen auszuwählen. Damit sollte auch klar sein, dass stets $\binom{n}{k} = \binom{n}{n-k}$ gilt.

(Wir benutzten tatsächlich Satz 3, um Satz 2 zu zeigen, aber Satz 3 ist auch in weiteren Situationen nützlich.)

Man wird leicht bemerken, dass für große Zahlen n solche Zahlen wie $\binom{n}{k}$ sehr groß werden können, z.B. auf keinen Taschenrechner mehr passen. Erst recht wird dann eine Frage nach der Wahrscheinlichkeit von *höchstens* k Treffern unbequem: Alle Werte $P(X = i), i = 0...n$ wären dann zu addieren. Gerade in diesen Fällen hilft die Näherung durch eine Normalverteilung. Aber auch für viel praktischerer Fragen sind die Normalverteilungen wichtig: Wie gut ist es, wenn man den unbekannten Mittelwert einer Größe durch ein arithmetisches Mittel nähert, das man in einer Stichprobe fand? So besprechen wir im übernächsten die Normalverteilungen selbst, anschließend die benötigte Technik zum Umgang mit Mittelwert und Streuung sowie praktische Anwendungen.

2.2. Hypergeometrische Verteilungen. Eng verwandt mit den Binomialverteilungen sind die sogenannten hypergeometrischen. Ähnlichkeit und Unterschied soll an entsprechenden Urnenmodellen veranschaulicht werden: Zieht man aus eine Urne n Kugeln, legt jedoch dabei die gezogene Kugel stets zurück, um erneut die Kugeln in der Urne zu mischen, so ist die Variable „Trefferzahl" (d.h. Anzahl der gezogenen Kugeln mit einer bestimmten Eigenschaft) binomialverteilt, mit den Parametern $p =$ Anteil der Trefferkugeln in der Urne, $n =$ Anzahl der gezogenen Kugeln. In diesem Falle kommt es offenbar nicht darauf an, wie viele Kugeln absolut in der Urne sind. Zieht man dagegen n Kugeln auf einmal heraus (oder einzeln, aber eben ohne Zurücklegen) und betrachtet wieder die Variable „Trefferzahl", so bemerkt man, dass die absolute Anzahl N der Kugeln in der Urne ein bedeutsamer Parameter ist. Wir leiten die Wahrscheinlichkeitsfunktion her, welche die Wahrscheinlichkeiten für die verschiedenen möglichen Trefferzahlen angibt: Es werden n Kugeln aus N gezogen, natürlich $n \leq N$. Das macht $\binom{N}{n}$ mögliche Fälle. Die günstigen Fälle für k Treffer haben folgende Anzahl: $\binom{K}{k} \cdot \binom{N-K}{n-k}$, wobei K die absolute Anzahl der „Trefferkugeln" in der Urne ist. Denn aus den Trefferkugeln sind genau k auszuwählen, aus den übrigen „Nietenkugeln" unabhängig $n-k$. Nach der Formel „Wahrscheinlichkeit = Anzahl der günstigen geteilt durch Anzahl der möglichen Fälle" (anwendbar bei lauter gleichwahrscheinlichen Fällen!) haben wir also:

Satz 4. *Sei X die Variable „Trefferzahl" beim Ziehen (ohne Zurücklegen) von n Kugeln aus einer Urne mit N Kugeln, davon K „Trefferkugeln". Dann heißt X $(N, K, n)-$ hypergeometrisch verteilt, und es gilt für $0 \leq n \leq N$, $0 \leq k \leq K$:*

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Vergleichen wir mit der Binomialverteilung mit demselben Wert n und $p = K/N$. Zunächst fällt auf, dass für $k > K$ die hypergeometrisch verteilte Trefferzahlgröße die Wahrscheinlichkeit für k Treffer gleich Null ist - für die entsprechende binomialverteilte gilt das nicht. Auch die andern Werte werden anders. Beispiel:

$N = 10$, $K = 5$, $n = 5$, $k = 2$. Binomialverteilung mit entsprechendem $p = 1/2$ ergibt die Wahrscheinlichkeit $\binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = 0.3125$, hypergeometrische Verteilung ergibt: $\binom{5}{2}\binom{5}{3}/\binom{10}{5} = 0.396$. Im Gegensatz dazu wird man finden, dass für N *viel* größer als n keine nennenswerten Unterschiede auftreten.

2.3. Eine Anwendung der hypergeometrischen Verteilungen: Fisher's exakter Test. Wir betrachten folgende Situation: Auf einer Population haben wir zwei Eigenschaften (sagen wir z.B. „Student" und „Internetbenutzer"). Das sind zwei Variablen mit jeweils zwei Ausprägungen „ja", „nein" bzw. „1", „0". Dann interessiert man sich häufig dafür, ob die eine mit der andern „irgendwie zu tun hat" oder nicht - man denke nicht an „Ursache - Wirkung"! Genauer (und viel allgemeiner als das leidige Schema von Ursache und Wirkung!) fragt man danach, ob die Verteilung der Ausprägungen der einen Eigenschaft dieselbe ist, gleichgültig, welche Ausprägung der andern Eigenschaft vorliegt. Dann nennt man die Merkmale „unabhängig" (im statistisch-wahrscheinlichkeitstheoretischen Sinne!), sonst „abhängig". Beispiele: Bei mehrmaligem Würfeln ist die nachfolgende Augenzahl unabhängig von allen vorangehenden, aber man erwartet, dass das Merkmal „Internetbenutzer" nicht unabhängig vom Merkmal „Student" ist, und zwar in dem Sinne, dass der Anteil der Internetnutzer bei den Studenten höher liegt als in der sonstigen Bevölkerung, *nicht in dem Sinne, dass jeder Student Internetnutzer wäre* (oder jeder Nichtstudent kein solcher)! Man sucht nun einen solchen Sachverhalt von Abhängigkeit empirisch zu ermitteln über eine Zufallsstichprobe. Im Beispiel möge man etwa gefunden haben (die Einträge bedeuten absolute Häufigkeiten):

	Internetnutzer	Kein Internetnutzer
Student	6	5
Kein Student	3	10

Wir setzen nun hypothetisch einmal voraus, diese Merkmale seien völlig unabhängig, und stellen uns das konkret so vor: Bei den 24 Leuten wurden 9 rein zufällig als „Internetbenutzer" ausgewählt, und von diesen 9 gerieten rein zufällig 6 in die Gruppe der 11 Studenten und 3 nur in die Gruppe der Nichtstudenten. Das könnte ja erst einmal so passiert sein: Wählen Sie 20 Leute zufällig aus, so werden Sie auch nicht genau 10 Frauen dabei haben! Aber, so fragen wir weiter: Ist es nicht sehr unwahrscheinlich, dass ein solch krasses Missverhältnis rein zufällig entsteht, wenn der Sachverhalt der Unabhängigkeit wie vorausgesetzt besteht? Genau diese Wahrscheinlichkeitsfrage beantworten wir mit Einsetzen der hypergeometrischen Verteilung, die genau unserem Unabhängigkeitsmodell entspricht; zunächst: Wie wahrscheinlich ist es, *genau die beobachtete* Tafel von Häufigkeiten zu erhalten? Aus 24 Leuten, davon 11 Studenten, wurden rein zufällig 9 ausgewählt als Internetnutzer, das macht $\binom{24}{9}$ gleichwahrscheinliche Fälle. Die Anzahl der günstigen Fälle für unsere Tafel ergibt sich daraus, dass unter den 9 Internetnutzern genau 6 Studenten „gezogen" wurden und 3 Nichtstudenten, das macht $\binom{11}{6}\binom{13}{3}$, also haben wir genau die Wahrscheinlichkeit

$$\frac{\binom{11}{6}\binom{13}{3}}{\binom{24}{9}} \approx 0.1$$

dafür, genau die obenstehende Tafel zu beobachten. Aber unsere Frage ist die nach der Wahrscheinlichkeit dafür, ein mindestens so starkes Missverhältnis zu beobachten, wir haben also die Wahrscheinlichkeiten aller Tafeln zu addieren, welche ein

solches darstellen. Es sind dies die folgenden:

(2.2)
$$\begin{array}{|c|c|}\hline 6&5\\\hline 3&10\\\hline\end{array},\ \begin{array}{|c|c|}\hline 7&4\\\hline 2&11\\\hline\end{array},\ \begin{array}{|c|c|}\hline 8&3\\\hline 1&12\\\hline\end{array},\ \begin{array}{|c|c|}\hline 9&2\\\hline 0&13\\\hline\end{array},$$
$$\begin{array}{|c|c|}\hline 2&9\\\hline 7&6\\\hline\end{array},\ \begin{array}{|c|c|}\hline 1&10\\\hline 8&5\\\hline\end{array},\ \begin{array}{|c|c|}\hline 0&11\\\hline 9&4\\\hline\end{array}.$$

Man beachte: Die zweite Reihe von Tafeln repräsentiert ein mindestens ebenso deutliches Missverhältnis wie beobachtet, nur nach der andern Seite: Unabhängigkeit bedeutet ja, dass auch nicht etwa unter den Nichtstudenten die Internetnutzer überrepräsentiert wären. Genauer ist einzusehen, dass man nicht etwa mit der Tafel

$$\begin{array}{|c|c|}\hline 3&8\\\hline 6&7\\\hline\end{array}$$

zu beginnen hätte: Bei dieser Tafel ist der Anteil der Internetnutzer unter den Studenten $3/11$, unter den Nichtstudenten $6/13$, die absolute Differenz zwischen diesen beiden Zahlen (die laut Hypothese als gleich *zu erwarten* wären), ist $6/13 - 3/11 = 27/143$. Bei der tatsächlich beobachteten Tafel haben wir eine Diskrepanz von $6/11 - 3/13 = 45/143$, und diese Zahl ist größer als $27/143$. Stellen Sie analog fest, dass dagegen die Tafel

$$\begin{array}{|c|c|}\hline 2&9\\\hline 7&6\\\hline\end{array}$$

ein größeres Missverhältnis als $45/143$ darstellt. Das Aufaddieren aller Wahrscheinlichkeiten für alle Tafeln aus 2.2 ergibt:

$$\frac{1}{\binom{24}{9}}\left(\binom{11}{6}\binom{13}{3}+\binom{11}{7}\binom{13}{2}+\binom{11}{8}\binom{13}{1}+\binom{11}{9}\binom{13}{0}\right.$$
$$\left.+\binom{11}{2}\binom{13}{7}+\binom{11}{1}\binom{13}{8}+\binom{11}{0}\binom{13}{9}\right),\ \text{das ist etwa } 0.21.$$

Nun zur Bewertung: Es ist nicht *sehr* unwahrscheinlich, bei reinem Zufall ein solches Missverhältnis zu erhalten wie beobachtet oder ein größeres. Wir haben also nicht gerade so etwas Seltenes wie einen Lotto-Hauptgewinn beobachtet. Damit ist die Hypothese nicht stark erschüttert, trotz des deutlichen Missverhältnisses in der Stichprobe. Das bedeutet aber keineswegs, dass wir die Hypothese glauben und als wahr annehmen sollten, vielmehr werden wir uns erklären: Die Stichprobe war recht klein, und größere Stichproben könnten die Hypothese sehr wohl noch zu Fall bringen. Eine Beobachtung etwa wie

$$\begin{array}{|c|c|}\hline 60&50\\\hline 30&100\\\hline\end{array}$$

zeigt dieselbe Diskrepanz, aber eine solche wäre *viel unwahrscheinlicher als* $1/5$ (das Resultat ist praktisch Null) unter der Hypothese, dass sie zufällig herauskäme, also die Merkmale in der gesamten Population unabhängig wären. Allerdings wäre es nur mittels eines Computerprogramms möglich, diese Wahrscheinlichkeit nach dem oben gezeigten Muster auszurechnen, weil es zu viele Summanden gibt und zudem die auftretenden Zahlen $\binom{n}{k}$ zu groß werden. Später lernen wir dazu den sogenannten χ^2- Test kennen, der gerade im Falle hoher Einträge eine sehr akkurate Näherung ergibt. Mittels des Computers kann man auch ohne weiteres

die Verallgemeinerung auf Merkmale mit mehr als zwei Ausprägungen noch nach obenstehendem Muster rechnen.

2.4. Die Normalverteilungen. Es handelt sich um die bekannte Glockenform (im Dichtebild), allerdings ist es eine ganz bestimmte Glockenform mit ihren durch bloßes Strecken, Stauchen und Verschieben gegebenen Modifikationen. (Mathematisch kann man eine unendliche Fülle völlig andersartiger Glockenkurven bilden.) Die folgende Definition benötigt man zum praktischen Umgang nicht, sie soll nur klarstellen, dass es sich um eine durch zwei Parameter definiterte feste Familie von Dichtefunktionen handelt. Außerdem werden die weiterhin oft zu benutzenden Bezeichnungen insbesondere für die zugehörigen Verteilungsfunktionen eingeführt.

DEFINITION 11. *Für jede Zahl $\mu \in \mathbb{R}$ und jede Zahl $\sigma > 0$ definiert man folgende Dichte:*

$$\boxed{\varphi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} \ (x \in \mathbb{R}))} \tag{2.3}$$

$\varphi_{\mu,\sigma}$ = Dichte der $(\mu,\sigma)-$Normalverteilung. Die zugehörige Verteilungsfunktion nennen wir $\Phi_{\mu,\sigma}$. Also:

$$\boxed{\begin{array}{ll}\Phi_{\mu,\sigma}(a) & = \text{Flächeninhalt unter dem Graphen der Dichte } \varphi_{\mu,\sigma} \text{ links von } a \\ & (\text{als Integral zu berechnen}) \ (\text{für alle } a \in \mathbb{R}).\end{array}}$$

Speziell für $\mu = 0, \sigma = 1$ ergibt sich die Standard-Normalverteilung, und man schreibet manchmal kurz: φ für $\varphi_{0,1}$, Φ für $\Phi_{0,1}$. Ist X eine $(\mu,\sigma)-$ normalverteilte Größe, so ist mit dieser Definition:

$$\boxed{P(X \le a) = \Phi_{\mu,\sigma}(a) = \int_{-\infty}^{a} \varphi_{\mu,\sigma}(x)dx \ (\text{für alle } a \in \mathbb{R}).} \tag{2.4}$$

Natürlich gilt mit Satz 1 wieder: $\Phi'_{\mu,\sigma} = \varphi_{\mu,\sigma}$. (Vgl. Abb. 5 für die Standard-Normalverteilung.)

Der folgende Satz 4 erklärt die überragende Bedeutung der Normalverteilungen:

SATZ 5.

Zentraler Grenzwertsatz (untechnische Formulierung):
Lange Summen unabhängiger Variablen sind annähernd normalverteilt, wenn nur die Streuungen der summierten Variablen in endlichen Grenzen und oberhalb einer festen Zahl > 0 bleiben, außerdem die dritten zentralen Momente (wie Varianzen gebildet, nur mit dritter Potenz) beschränkt bleiben.

Insbesondere sind all diese Bedingungen erfüllt im praktischen Hauptfall der Anwendung, dass es sich bei den Summanden um unabhängige Kopien ein und derselben Variable handelt, wie man bei Stichproben hat (genau genommen müsste man sie „mit Zurücklegen" ziehen, um Unabhängigkeit zu wahren, aber das kann man getrost verletzen, wenn der Umfang gering gegen den Populationsumfang ist).

Bemerkung: Was „hinreichend lang" ist, muss man aus Erfahrung lernen. Faustregel für nicht allzu anti-normalverteilte zu summierende Variablen (also nicht extrem schief oder U-förmig): Länge 10 ist schon ziemlich gut. Um einen Eindruck davon zu geben, betrachten wir folgende Abbildung 8, welche die Entwicklung von einer Gleichverteilung auf dem Intervall $[-1, 1]$ zur Normalverteilung durch Summenbildung unabhängiger Einzelwerte für verschiedene Summenlängen zeigt. Damit die Verteilungen nicht immer breiter werden, dividieren wir dabei die Summen der Länge n stets durch n, zeigen also jeweils die Verteilung der Variablen: Arithmetisches Mittel von n zufällig ausgewählten Werten einer auf $[-1, 1]$ gleichverteilten Größe.

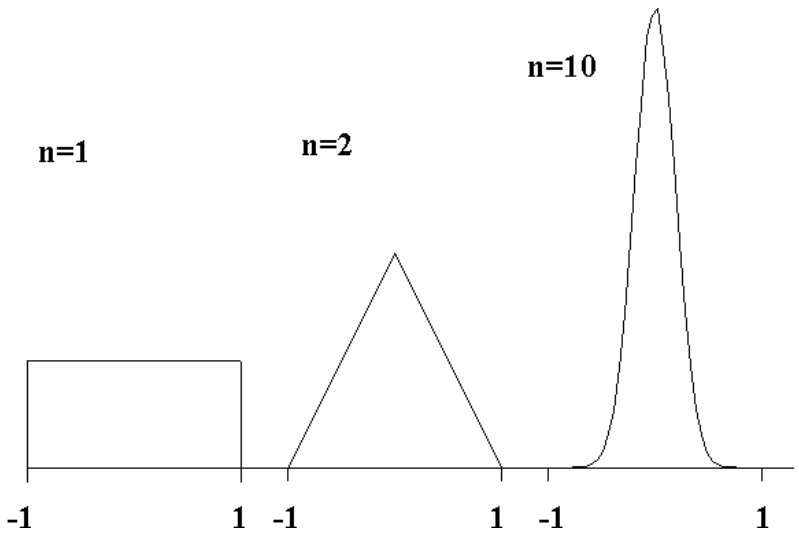


Abb. 8: Die sukzessive Näherung an die Normalverteilung durch Summenbildung unabhängiger Variablen, hier z.B. gleichverteilter Variablen auf $[-1, 1]$ - n gibt die Zahl der unabhängigen Summanden.

Hier sind drei weitere Illustrationen des Zentralen Grenzwertsatzes: Binomialverteilte Variablen sind für große n lange Summen unabhängiger Kopien ein und derselben Bernoulli-Variablen. Somit sollte bei p, das nicht zu weit von $1/2$ liegt, sehr schnell, sonst (wegen der zunächst starken Asymmetrie) langsamer, also erst für größere n, eine Normalverteilung angenähert erscheinen.

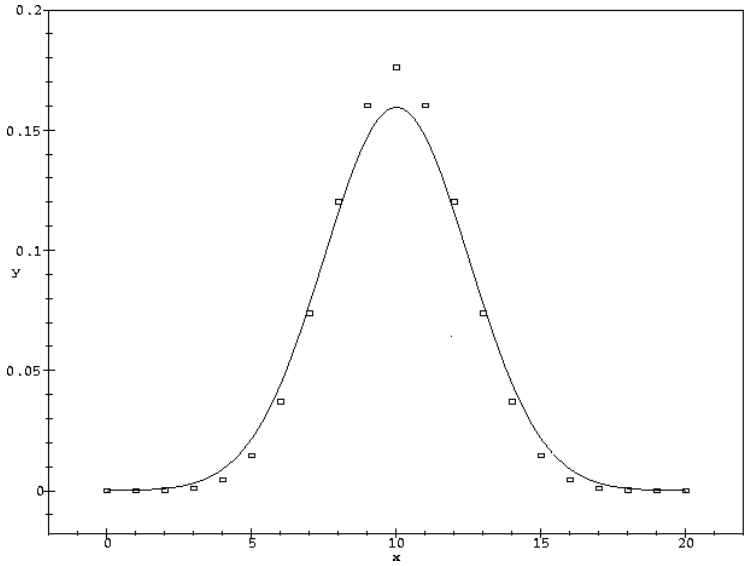


Abb. 9: Bereits recht gute Annäherung an eine Normalverteilung durch die Binomialverteilung mit $n = 20$, $p = 1/2$ (die Quadrate geben die Werte der Wahrscheinlichkeitsfunktion)

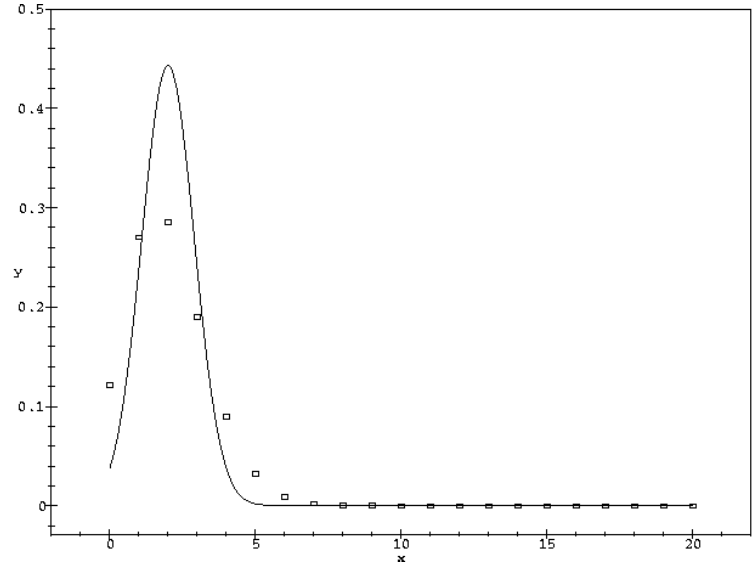


Abb. 10: Weniger gute Annäherung an eine Normalverteilung mit der schiefen Binomialverteilung zu $n = 20$, $p = 1/10$

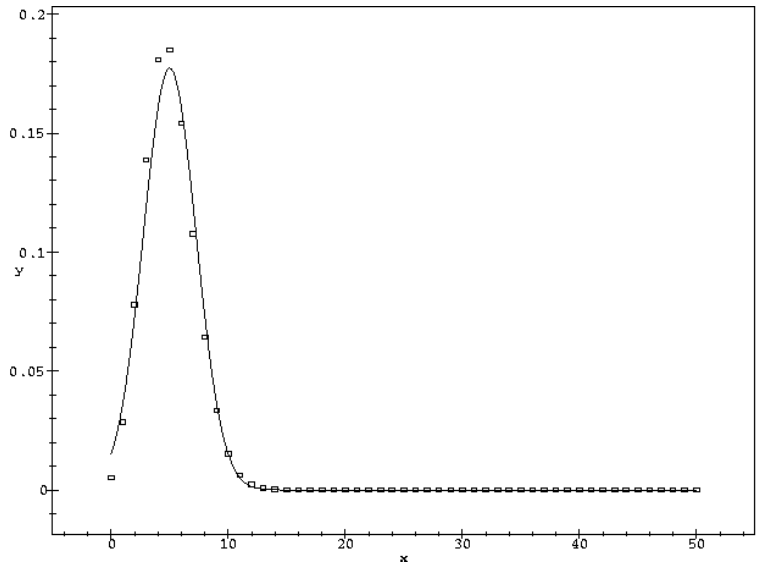


Abb. 11: Auch bei schiefer Binomialverteilung gute Annäherung an eine Normalverteilung mit wachsendem n, hier $n = 50$, wieder $p = 1/10$.

Dabei ist natürlich zu beachten, dass man immer nur eine ordentliche Annäherung an die jeweils *passende* Normalverteilung erhält, d.h. diejenige mit den richtigen Werten μ, σ. (Vgl. den nächsten Abschnitt 3. für das Rechnen mit μ, σ.) Auch die hypergeometrischen Verteilungen nähern sich Normalverteilungen an, wenn nur N recht viel größer als n ist und n ebenfalls relativ zum Abstand von K/N zu $1/2$ nicht zu klein. Hier zwei Beispiele:

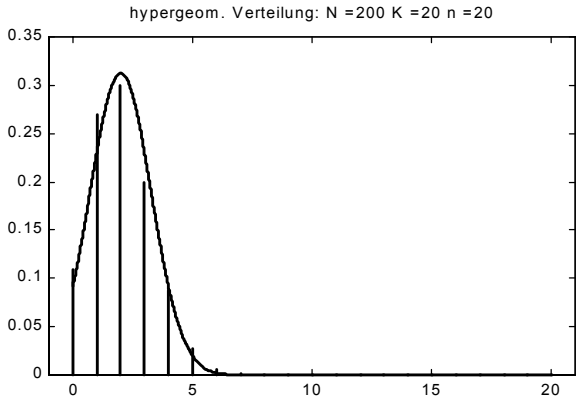


Abb. 12: Annäherung an die entsprechende Normalverteilung ist schon nicht schlecht bei der hypergeometrischen Verteilung zu $N = 200$, $K = 20$, $n = 20$.

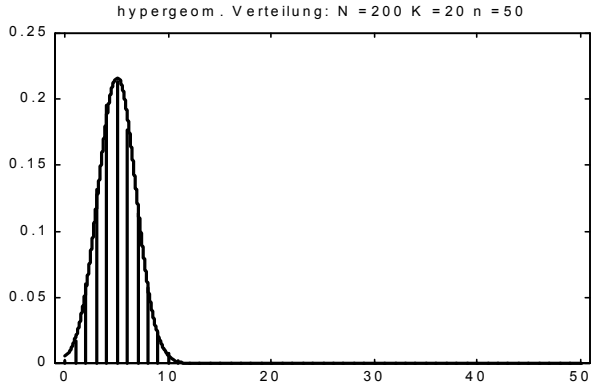


Abb. 13: Für $n = 50$ (weiterhin $N = 200$, $K = 20$) wird die Annäherung an die entsprechende Normalverteilung auch für die hypergeometrische deutlich besser.

Zwei Umstände machen den Zentralen Grenzwertsatz so interessant für mannigfachste Anwendungen: Einmal erklärt er, warum man so viele Normalverteilungen als „natürliche" in der Welt antrifft: Immer dann, wenn sich unabhängig gewisse Variablen mitteln, um einen Wert zu bestimmen (Erbschaften, Ernährung usw. zur Bestimmung der Körperlänge eines Erwachsenen; Erbschaften und Anregungen und mehr zur Entwicklung einer Intelligenz), dann kristallisiert sich eine Normalverteilung heraus. Zweitens: Einige wiederkehrend zu benutzende Verteilungen von großer praktischer Bedeutung erweisen sich sofort als annähernde Normalverteilungen. Vor allem sind Mittelungsgrößen zu nennen: Man geht aus von einer Größe X, betrachtet nun aber Stichproben aus deren Population Ω, Stichproben eines festen Umfang n. So entsteht eine neue Größe, indem man jeder Stichprobe ihr arithmetisches Mittel der $X-$ Werte zuordnet. Diese Variable nennen wir $\overline{X}$. Man beachte, dass der Stichprobenumfang in dieser Notation unterdrückt ist. Man teilt ihn gewöhnlich in Worten oder Zusätzen wie „$n = ...$" mit. Einen beobachteten Wert von $\overline{X}$ bezeichnet man konsequent mit $\overline{x}$ und benutzt ihn als Schätzwert für den meist unbekannten Wert $\mu(X)$. Gewöhnlich kennt man ja nicht sämtliche X-Werte der Gesamtpopulation. Nun wissen wir aber mit dem Zentralen Grenzwertsatz, dass für hinreichend große n die Variable $\overline{X}$ annähernd normalverteilt ist (für das arithmetische Mittel bildet man eine Summe unabhängiger Einzelwerte, und die anschließende Division durch n ergibt nur ein Stauchen/Strecken im Verteilungsbild, ändert daher nichts am Normalverteilungscharakter, wie auch im Beispiel illustriert), und somit werden wir mit der praktischen Beherrschung der Normalverteilungen aussagen können, mit welcher Sicherheit (oder Wahrscheinlichkeit) bei einer solchen Schätzung der Fehler höchstens wie groß ist. Wir können uns also von der Qualität der Schätzung ein *quantitativ genaues* Bild machen, *ohne μ zu kennen!* Auf diese Weise führt uns die Normalverteilung auch in die „Schließende Statistik" (auch „Inferenzstatistik" genannt).

3. Bedingte Wahrscheinlichkeiten und Unabhängigkeit

Es gibt verschiedene Begriffe von „Unabhängigkeit". In unserem Kontext kommt es vor allem darauf an, den wahrscheinlichkeitstheoretischen Begriff der Unabhängigkeit richtig zu verstehen und insbesondere nicht mit dem logischen zu verwechseln. Eine Aussage A hängt logisch von einer andern Aussage B ab, wenn

man aus B entweder A oder aber „*nicht* A" schließen kann. Zum Beispiel: Wenn der erste Wurf mit einem Würfel eine Drei ergibt (A), dann muss die Augensumme nach zwei Würfen einen Wert unter 10 ergeben (B). B ist also logisch abhängig von A. Abhängigkeit im wahrscheinlichkeitstheoretischen Sinne ist dagegen eine schwächere Beziehung, man kann sie auch als Verallgemeinerung logischer Abhängigkeit auffassen: Die Ereignisse A („erster Wurf eine Drei") und C („Augensumme nach 2 Würfen unter 9") sind nicht logisch abhängig: Aus A kann man weder C noch „*nicht* C" logisch folgern. Aber unter der Bedingung A ist C wahrscheinlicher als C ohne Voraussetzung irgendeiner Bedingung wäre; dies bedeutet genau: C ist im wahrscheinlichkeitstheoretischen Sinne abhängig von A. Diesen Begriff wollen wir nun genauer quantitativ klären. Dazu definieren wir:

DEFINITION 12. *Für* $P(B) \neq 0$ *wird definiert:*

$$P(A\,|\,B) := \frac{P(A \cap B)}{P(B)} \text{ (lies: Wahrscheinlichkeit von } A \text{ bedingt durch } B).$$

Zur inhaltlichen Interpretation: Man schaut B als eingeschränkte Menge von möglichen Ausgängen an und bestimmt in diesem Rahmen die Wahrscheinlichkeit von A. Selbstverständlich sind nunmehr die „günstigen Fälle" diejenigen in $A \cap B$. Auf keinen Fall interpretiere man B als „Ursache von A" oder auch nur als zeitlich auf A folgend.

Wichtige Bemerkung zur Definition: Man verwendet diese Definition fast nie, um eine bedingte Wahrscheinlichkeit zu berechnen, vielmehr ergeben sich die Werte bedingter Wahrscheinlichkeiten meist unmittelbar. Dagegen ist sie von theoretischer Bedeutung für weitere Zusammenhänge und Formeln (s.u.). Die Division durch $P(B)$ ergibt sich gerade daraus, dass B als neue Menge Ω zu betrachten ist, und es sollte $P(B\,|\,B) = 1$ sein.

Nun können wir vorläufig A und B *im wahrscheinlichkeitstheoretischen Sinne unabhängig* nennen, wenn $P(A\,|\,B) = P(A)$, falls $P(B) \neq 0$. Beispiel: Zwei mal wird gewürfelt. A : „Erster Wurf eine Drei", B : „Zweiter Wurf eine Vier". Offenbar ist $P(A\,|\,B) = 1/6 = P(A)$, also A und B unabhängig. (Man beachte: B folgt sogar zeitlich auf A,was dem Sinn von $P(A\,|\,B)$ keinen Abbruch tut. Ausführliche Berechnung von $P(A\,|\,B)$: Im verkleinerten Topf B sind die Ausgänge $(1,4),(2,4),(3,4),(4,4),(5,4),(6,4)$. Das sind 6 gleichwahrscheinliche. In $A \cap B$ ist nur $(3,4)$, also ein günstiger Fall, macht Wahrscheinlichkeit $1/6$. Dagegen sind B und das folgende Ereignis C abhängig: C : „Die Augensumme beider Würfe ist 5". Denn offenbar $P(C\,|\,B) = 1/6$, während $P(C) = 4/36 = 1/9$. (Man zähle das nach.)

Wir kommen zur theoretischen Bedeutung der bedingten Wahrscheinlichkeiten:

SATZ 6. *Es gilt allgemein für* $P(B) \neq 0$:

$$P(A \cap B) = P(A\,|\,B)P(B). \tag{3.1}$$

Dazu ist nur die definierende Gleichung für $P(A\,|\,B)$ mit $P(B)$ zu multiplizieren. Diese Formel ist so häufig direkt anwendbar wie man direkt an $P(A\,|\,B)$ herankommt. Beispiel: Man zieht aus einer Urne mit 10 Kugeln, davon 5 rot, zwei Kugeln ohne Zurücklegen. Sei A das Ereignis: „Die zweite Kugel ist rot", B das Ereignis: „Die erste Kugel ist rot". Dann ist $P(A \cap B) = \frac{4}{10} \cdot \frac{5}{10} = \frac{1}{5}$; denn nach dem Herausziehen einer roten Kugel verbleiben 9 Kugeln, davon 4 rote, so dass die Wahrscheinlichkeit für das Ziehen einer zweiten roten Kugel 4/10

beträgt. Weiter können wir folgern, dass bei unabhängigen Ereignissen A, B stets $P(A \cap B) = P(A)P(B)$ ist, da $P(A\,|B) = P(A)$ in diesem Falle gilt. Um nun die Randfälle $P(B) = 0$ noch mitzunehmen, in denen diese Formel ebenfalls gilt, *definiert* man:

DEFINITION 13. *Zwei Ereignisse A und B (im Rahmen eines Zufallsexperiments) heißen unabhängig, wenn*

$$P(A \cap B) = P(A)P(B).$$

(Man beachte, dass dies für $P(B) \neq 0$ gleichwertig zu $P(A\,|B) = P(A)$ ist.)

Eine sehr wichtige Verallgemeinerung dieses Begriffs auf Variablen kann man daraus entwickeln: Die für eine Variable X interessierenden Ereignisse lauten $X \leq a$, $a \in \mathbb{R}$. Denn deren Wahrscheinlichkeiten bestimmen bereits die Verteilungsfunktion. Somit kann man in typisch mathematischer Weise den Begriff der Unabhängigkeit von Variablen auf den von Ereignissen zurückführen:

DEFINITION 14. *Zwei Variablen X, Y heißen unabhängig, wenn für jedes Paar (a, b) reeller Zahlen gilt: Die Ereignisse $X \leq a$ und $Y \leq b$ sind unabhängig.*

Beispiel: Körperlänge und Körpergewicht (auf der Population der Menschen) sind sicher nicht unabhängig; zwar gibt es kurze Dicke und lange Dünne, aber wenn die Körperlänge unter einer Grenze wie 160 cm bleibt, so ist jedenfalls die Wahrscheinlichkeit für ein Körpergewicht unter 60 kg erhöht, d.h. größer als unter allen Leuten. Ein wichtiger Teil der Statistik (vgl. das letzte Kapitel) beschäftigt sich damit, die Abhängigkeiten zwischen mehreren Variablen zu beschreiben.

Wir kommen zu zwei weiteren recht wichtigen Formeln der Wahrscheinlichkeitsrechnung, die mit bedingten Wahrscheinlichkeiten arbeiten:

SATZ 7 (Bayessche Formel und Formel von der totalen Wahrscheinlichkeit). *Es seien $P(A), P(B) \neq 0$. Dann gilt folgende Bayessche Formel:*

$$P(B\,|A) = \frac{P(A\,|B)P(B)}{P(A)} \tag{3.2}$$

Sei $\{B_1, ..., B_n\}$ eine Klasseneinteilung von Ω, d.h.

$$\Omega = \bigcup_{i=1}^{n} B_i \text{ und } B_i \cap B_j = \emptyset \text{ für } i \neq j,\ 1 \leq i, j \leq n.$$

Dann gilt folgende Formel von der totalen Wahrscheinlichkeit:

$$P(A) = \sum_{i=1}^{n} P(A\,|B_i)P(B_i). \tag{3.3}$$

Die Formeln sind ungeachtet ihrer Nützlichkeit sehr einfach zu zeigen: Die erste ergibt sich sofort aus $P(A \cap B) = P(A\,|B)P(B)$ und $P(B \cap A) = P(B\,|A)P(A)$ nach vorigem Satz, aber $A \cap B = B \cap A$. Damit $P(A\,|B)P(B) = P(B\,|A)P(A)$, nun teile man diese Gleichung durch $P(B)$. Der Nutzen der Formel besteht darin, dass man bedingte Wahrscheinlichkeiten „herumdrehen" kann: Man stößt vielfach auf Situationen, in denen $P(B\,|A)$ inhaltlich interessiert, aber nut $P(A\,|B)$ empirisch zugänglich ist. Etwa A : Auftreten eines Symptoms, B : Vorliegen einer bestimmten Krankheit. Diagnostisch wichtig ist $P(B\,|A)$, aber nur $P(A\,|B)$ ist empirisch zugänglich. Ebenso sind $P(A\,|\overline{B})$ sowie $P(B)$ zugänglich, und so kann man mittels der zweiten Formel den Nenner ausrechnen als $P(A) = P(A\,|B)P(B) + P(A\,|\overline{B})P(\overline{B})$.

Mit der Bayesschen Formel hat man dann $P(B\,|A\,)$, also im Beispiel die Wahrscheinlichkeit dafür, dass jemand die Krankheit hat, bei dem das Symptom auftritt. Es gibt noch einen theoretisch weiter reichenden Zweig von Bayesscher Statistik, der eine recht große Bedeutung erlangt hat und insbesondere eine weiterführende Art des induktiven Schließens aus Empirischem darstellt, und auch hier steht dies Herumdrehen bedingter Wahrscheinlichkeiten am Anfang der Überlegungen: Stellen Sie sich vor, Sie haben verschiedene Theorien $T_1, ...T_n$, sagen wir Wahrscheinlichkeitsmodelle, als mögliche Erklärungen für einen Phänomenbestand, und fragen sich, welche dieser Theorien am besten passt zu allen Beobachtungsdaten B (als ein komplexes Ereignis aufgefasst). Nun können Sie $P(B\,|T_i\,)$ jeweils bestimmen, d.h. die Wahrscheinlichkeit für die Beobachtungen bei Voraussetzung der Theorie T_i, $1 \leq i \leq n$. Ein primitiveres Prinzip (man nennt es „Maximum Likelihood") würde nun einfach eine solche Theorie bevorzugen, bei der B maximale Wahrscheinlichkeit erhält. Das elaboriertere Bayesprinzip lautet stattdessen: Man fasse die Gültigkeit einer Theorie wiederum als ein Ereignis auf, schreibe also den Theorien selbst Wahrscheinlichkeiten zu, und zwar bedingte durch die Beobachtung, gemäß Bayesscher Formel:

$$P(T_i\,|B\,) = \frac{P(B\,|T_i\,)P(T_i)}{P(B)}.$$

Interessant sind hier die Ausgangswahrscheinlichkeiten $P(T_i)$ für die Theorien; setzt man sie alle gleich, so landet man wieder bei Maximum Likelihood: $P(T_i\,|B\,)$ wird am höchsten für die Theorien, für die $P(B\,|T_i\,)$ maximal ist. Aber man kann auch Vorerfahrungen oder plausible Annahmen einfließen lassen, um diese Ausgangswahrscheinlichkeiten differenzierter anzusetzen.

Zur Begründung der Formel von der totalen Wahrscheinlichkeit ist nur zu bemerken:

$$P(A) = P\,(\Omega \cap A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A\,|B_i\,)P(B_i)$$

Dabei wurde benutzt, dass $\Omega \cap A = \bigcup_{i=1}^{n} (A \cap B_i)$ und die Summenformel 1.1 aus den Axiomen sowie Formel 3.1. Die Formel 3.3 ist nicht etwa nur im Bayes-Zusammenhang zu nutzen, sondern elementar tritt häufig die Situation auf, dass man bedingte Wahrscheinlichkeiten zur Verfügung hat und eine unbedingte sachlich interessiert: Zum Beispiel ist die relative Häufigkeit einer Eigenschaft bei allen Bundesbürgern auf diese Weise zu ermitteln, wenn man die relativen Häufigkeiten für die einzelnen Bundesländer und dazu die Bevölkerungsanteile der Bundesländer an der gesamten Republik kennt.

4. Das Rechnen mit μ und σ

Es gibt einige sehr nützliche Formeln für das Rechnen mit μ und σ, die insbesondere im Zusammenhang mit der Normalverteilung sehr wichtig sind - erinnern wir uns daran, dass Kenntnis von μ, σ bei einer Normalverteilung ausreicht, um jegliche Wahrscheinlichkeitsfrage zu beantworten. Wir stellen diese Formeln in zwei Blöcken zusammen - der erste umfasst nur stets gültige Formeln, der zweite solche, die gewissen Voraussetzungen unterliegen.

SATZ 8. *Es gelten stets folgende Formeln - seien X, Y Variablen auf derselben Population, $a \in \mathbb{R}$:*

$$\left.\begin{array}{c}\mu(aX) = a\mu(X) \\ \mu(X+Y) = \mu(X) + \mu(Y)\end{array}\right\} \text{ Linearität von } \mu \tag{4.1}$$

$$\begin{aligned}\sigma^2(aX) &= a^2\sigma^2(X),\ \sigma(aX) = |a|\,\sigma(X),\ \sigma(X+a) = \sigma(X) \\ \sigma^2(X) &= \mu\left(X^2\right) - \mu^2(X).\end{aligned}$$

Eine wichtige Folgerung: Für $\sigma(X) \neq 0$ hat man

$$\begin{aligned}\mu\left(\frac{X-\mu(X)}{\sigma(X)}\right) &= 0 \\ \sigma\left(\frac{X-\mu(X)}{\sigma(X)}\right) &= 1.\end{aligned}$$

Man kann also jede Variable mit nichtverschwindender Varianz durch diese lineare Transformation der Standardisierung auf Mittelwert Null und Varianz (damit auch Streuung) 1 bringen.

Alle diese Formeln sind sehr leicht nachzurechnen. Die Formel $\sigma^2(X) = \mu\left(X^2\right) - \mu^2(X)$ wird weiter unten allgemeiner für die Kovarianz bewiesen.

Zunächst ist noch ein Spezialfall von Abhängigkeit zweier Variablen zu definieren, damit die Voraussetzungen für weitere Formeln angemessen formuliert werde können. Die zugehörigen Überlegungen ergeben eine weitere nützliche Formel für die Varianz.

Rechnet man $\sigma^2(X+Y)$ aus, so erhält man

$$\begin{aligned}\sigma^2(X+Y) &= \mu\left((X+Y)^2\right) - \mu^2(X+Y) \\ &= \mu\left(X^2\right) + \mu\left(Y^2\right) + 2\mu(XY) - \mu^2(X) - \mu^2(Y) - 2\mu(X)\mu(Y) \\ &= \sigma^2(X) + \sigma^2(Y) + 2\left(\mu(XY) - \mu(X)\mu(Y)\right).\end{aligned}$$

Man sieht daran, dass sich im allgemeinen die Varianzen der Variablen nicht zur Varianz der Summe addieren, sondern ein Zusatzterm entsteht. Dieser wird noch im Zusammenhang mit linearer Regression interessieren, und daher heben wir ihn mit einer gesonderten Definition hervor:

DEFINITION 15. *Die Kovarianz der Variablen X, Y, welche auf derselben Population Ω definiert seien, ist definiert als*

$$\begin{aligned}Cov(X,Y) \;:\; &= \mu\left((X-\mu(X))(Y-\mu(Y))\right). \text{ Es folgt:} \\ &\text{Man schreibt auch } \sigma^2(X,Y) \text{ für } Cov(X,Y).\end{aligned}$$

Man beachte die Analogie zur Varianz: Setzt man $X = Y$, so ist $Cov(X,X) = \sigma^2(X)$. Allerdings kann eine Kovarianz bei verschiedenen Variablen durchaus negativ sein. Der folgende einfache Satz zeigt die wichtigen Rechengesetze für die Kovarianz:

SATZ 9. *Es gelten folgende Formeln für die Kovarianz - seien X, Y, Z Variablen auf derselben Population und a eine beliebige reelle Zahl:*

(4.2)

$$\begin{array}{lcl} (i) & \left.\begin{array}{c} Cov(aX,Y) = aCov(X,Y) \\ Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z) \end{array}\right\} & \textit{Linearität der Kovarianz im ersten Schlitz, macht Bilinearität mit (ii)} \\ (ii) & Cov(X,Y) = Cov(Y,X) & \\ (iii) & Cov(X,Y) = \mu(XY) - \mu(X)\mu(Y). & \end{array}$$

Die Gleichungen (i) folgen sofort aus Kommutativgesetzen für $+, \cdot$ und dem Distributivgesetz für die reellen Zahlen sowie der Linearität von μ. Zeigen wir exemplarisch die zweite Gleichung von (i), die sich bei Gebrauch eben auch wie ein Distributivgesetz „anfühlt":

$$\begin{array}{rcl} Cov(X+Y,Z) & = & \mu\left[(X+Y-\mu(X+Y))\cdot(Z-\mu(Z))\right] \\ & = & \mu\left[(X-\mu(X))\cdot(Z-\mu(Z)) + (Y-\mu(Y)\cdot(Z-\mu(Z))\right] \\ (\text{Linearität von } \mu) & = & \mu\left[(X-\mu(X))\,(Z-\mu(Z))\right] + \mu\left[(Y-\mu(Y)\,(Z-\mu(Z))\right] \\ & = & Cov(X,Z) + Cov(Y,Z) \end{array}$$

Die Gleichung (ii) folgt sofort aus dem Kommutativgesetz für die Multiplikation reeller Zahlen. Man beachte, dass (i) und (ii) zusammen auch ergeben, dass $\sigma^2(X, aY) = a\sigma^2(X,Y)$ und $\sigma^2(X, Y+Z) = \sigma^2(X,Z) + \sigma^2(Y,Z)$, d.h. die Linearität von σ^2 im 2. Schlitz, was man dann zusammen Bilinearität nennt.

Die Gleichung (iii) kann man so beweisen (damit haben wir insbesondere auch die oben angeführte entsprechende Formel für die Varianz bewiesen, da diese nur den Spezialfall $X = Y$ darstellt):

$$\begin{array}{rcl} \mu\left((X-\mu(X))\,(Y-\mu(Y))\right) & = & \mu\left(XY - X\mu(Y) - Y\mu(X) + \mu(X)\mu(Y)\right) \\ & = & \mu(XY) - 2\mu(X)\mu(Y) + \mu(X)\mu(Y) \\ & = & \mu(XY) - \mu(X)\mu(Y). \end{array}$$

Die Sache funktioniert also mit einfachem Ausrechnen der zusammengesetzten Variablen und anschließender Nutzung der Linearität von μ.

Wir haben oben gesehen, dass die naiv zu erwartende Formel $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$ genau dann gilt, wenn $Cov(X,Y) = 0$, wenn also der Zusatzterm verschwindet. Später (im Abschnitt über lineare Regression) werden wir genauer verstehen, dass dies gleichbedeutend mit *linearer Unabhängigkeit* von X, Y ist und was diese *lineare* Unabhängigkeit bedeutet. Man achte genau darauf, dass der mathematische Sprachgebrauch hier einfach rein sprachlich ordentlich ist: Unabhängigkeit der Variablen impliziert lineare Unabhängigkeit, aber nicht umgekehrt. Lineare Abhängigkeit dagegen impliziert Abhängigkeit überhaupt, aber nicht umgekehrt. Nunmehr *definieren* wir einfach nur:

DEFINITION 16. *Zwei Variablen X, Y heißen linear unabhängig, wenn ihre Kovarianz den Wert Null hat, also $Cov(X,Y) = 0$ gilt.*

Wir haben den folgenden einfachen Satz, der besagt, dass Unabhängigkeit (überhaupt) die lineare Unabhängigkeit impliziert:

SATZ 10. *Sind X, Y unabhängig, so sind X, Y auch linear unabhängig, also $Cov(X,Y) = 0$.*

Zum Beweis ist zu zeigen: Unter der Voraussetzung der Unabhängigkeit gilt $\mu(XY) = \mu(X)\mu(Y)$. Da wir nur für Variablen X mit nur endlich vielen Werten $a \in \mathbb{R}$, an denen $f_X(a)$ nicht verschwindet, $\mu(X)$ definiert haben, können wir die Sache natürlich auch nur für diesen Fall beweisen, sie ist aber allgemeiner gültig. Wir haben unter der genannten Voraussetzung für X und Y:

$$\begin{aligned}
\mu(XY) &= \sum_{a,b\in\mathbb{R}} abP(X=a \text{ und } Y=b) \\
\text{(Unabhängigkeit von } X,Y) &= \sum_{a,b\in\mathbb{R}} abP(X=a)P(Y=b) \\
\text{(Distributivgesetz, Rechnen mit } \Sigma) &= \sum_{a\in\mathbb{R}} aP(X=a)\sum_{b\in\mathbb{R}} bP(Y=b) \\
&= \mu(X)\mu(Y)
\end{aligned}$$

Damit folgen mit den vorangehenden Argumenten sofort die weiterführenden Formeln für μ, σ:

SATZ 11. *Alle folgenden Formeln gelten* ***unter Voraussetzung der linearen Unabhängigkeit*** *von X, Y, also insbesondere auch dann, wenn X, Y überhaupt unabhängig sind:*

$$\begin{array}{ll}
(i) & \mu(XY) = \mu(X)\mu(Y) \\
(ii) & \sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) \\
(iii) & \sigma(X+Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)}
\end{array} \tag{4.3}$$

Wiederholtes Anwenden von (ii) ergibt die wichtige verallgemeinerte Formel: Sind $X_1, ..., X_n$ linear unabhängig (und dafür genügt bereits die paarweise lineare Unabhängigkeit dieser Variablen, was nicht etwa rein logisch selbstverständlich ist, aber eben für die ***lineare*** *Unabhängigkeit von Variablen - übrigens nicht für die Unabhängigkeit schlechthin! - zutrifft), so hat man*

$$\begin{aligned}
(iv) \quad \sigma^2\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n \sigma^2(X_i) \text{ und daher} \\
(v) \quad \sigma\left(\sum_{i=1}^n X_i\right) &= \sqrt{\sum_{i=1}^n \sigma^2(X_i)}
\end{aligned} \tag{4.4}$$

Man vermeide Stümpereien wie die Anwendung dieser Formeln, ohne dass die entscheidende Voraussetzung erfüllt ist, oder wie das lineare Rechnen mit der Wurzel - es ist $\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$, wenn nur a, b beide von Null verschieden sind.

Abschließend wollen wir noch die bereitgestellten Formeln dazu nutzen, Mittelwert und Varianz für binomialverteilte und hypergeometrisch verteilte sowie für Stichprobenmittelgrößen herauszupräparieren. Das ist insbesondere dafür nützlich, auf die betreffenden Variablen dann die Normalverteilungen anwenden zu können, für die man bekanntlich gerade μ, σ wissen muss.

Vorab sei noch einmal daran erinnert, dass zu einer Varialen X mit $\overline{X}$ bei vorausgesetztem Stichprobenumfang n (in der Notation tritt er nicht auf) die folgende Größe gemeint ist: *Jeder Stichprobe* aus Ω (wobei Ω die Population zu X ist) wird das arithmetische Mittel der darin vorzufindenden $X-$ Werte zugeordnet. Man beachte also: Die Population zu $\overline{X}$ ist die Menge aller Teilmengen aus Ω vom

Umfang n! Wir fassen die angesprochenen nützlichen Resultate alle zu einem Block zusammen:

SATZ 12. *(i) Sei X $(n,p)-$ binomialverteilt. Dann gilt:*

$$\begin{aligned}\mu(X) &= np\\ \sigma^2(X) &= np(1-p).\end{aligned}$$

(ii) Sei X $(N,K,n)-$ hypergeometrisch verteilt und $N>1$. Dann gilt:

$$\begin{aligned}\mu(X) &= n\frac{K}{N}\\ \sigma^2(X) &= n\frac{K}{N}\cdot\frac{N-K}{N}\cdot\frac{N-n}{N-1}\end{aligned}$$

(iii) Sei X in beliebiger Weise verteilt. Dann gilt für die Stichprobenmittelgröße $\overline{X}$ zum Stichprobenumfang $n\geq 1$:

$$\begin{aligned}\mu\left(\overline{X}\right) &= \mu(X)\\ \sigma\left(\overline{X}\right) &= \frac{\sigma(X)}{\sqrt{n}},\ \textit{entsprechend } \sigma^2\left(\overline{X}\right)=\frac{\sigma^2(X)}{n}.\end{aligned}$$

Zur Begründung von (i): Man rechnet einfach aus, dass für $n=1$ herauskommt: $\mu=p$ und $\sigma^2=p(1-p)$. Nunmehr fasst man eine $(n,p)-$ binomialverteilte Variable X als Summe von n unabhängigen $(1,p)-$ binomialverteilten (bzw. $p-$ Bernoulliverteilten) Variablen X_i auf. Dann ist $\mu(X)=\mu(\sum_{i=1}^n X_i)=\sum_{i=1}^n\mu(X_i)=np$, da $\mu(X_i)=p$ für $1\leq i\leq n$. Ebenso wird die Varianz (wegen der Unabhängigkeit, also insbesondere linearen Unabhängigkeit der X_i) n mal so groß, mit Formel 4.4. Völlig analog läuft die Begründung für (iii); denn man hat $\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i$, wobei jede der Variablen X_i dieselbe Verteilung wie X hat und zudem diese Variablen unabhängig sind. Also hat man

$$\begin{aligned}\mu\left(\overline{X}\right) &= \frac{1}{n}\sum_{i=1}^n\mu(X_i)=\frac{1}{n}n\mu(X)=\mu(X)\ \text{und}\\ \sigma^2\left(\overline{X}\right) &= \frac{1}{n^2}\sum_{i=1}^n\sigma^2(X_i)=\frac{1}{n^2}n\sigma^2(X)=\frac{\sigma^2(X)}{n}.\end{aligned}$$

(Für das Herausziehen eines Faktors $1/n^2$ bei der Varianz vgl. 4.1.)

Verbleibt noch die Berechnung für die hypergeometrische Verteilung. Sie ist für die Varianz ein wenig schwieriger, da keine Unabhängigkeit (auch keine lineare) vorliegt. Wir begnügen uns zunächst für μ mit der Bemerkung, dass die Formel für μ unter (ii) genau der unter (i) entspricht, da K/N eben dem Parameter p bei der Binomialverteilung entspricht. Weiter sehen wir die Ähnlichkeit bei σ^2: Der Faktor $(N-K)/N$ entspricht gerade dem Faktor $1-p$ bei der Varianz der Binomialverteilung. Nun kommt ein neuer hinzu, $(N-n)/(N-1)$. Offenbar ist der vernachlässigbar (wegen eines Wertes ≈ 1), wenn n viel kleiner als N ist. Das sollte einleuchten, da die hypergeometrische Verteilung sich dann der entsprechenden Binomialverteilung annähert. Wenn aber n nicht winzig gegen N ist, so sollte plausibel sein, dass man eine geringere Streuung bekommt als bei der entsprechenden Binomialverteilung, denken wir uns insbesondere den Extremfall $n=N$; dann werden alle Kugeln aus der Urne gezogen, und wir erhalten stets K als Trefferkugelzahl. Genau dies zeigt die Varianzformel, es kommt in diesem Falle Varianz Null

heraus. In derartigen Fällen liegt natürlich keine Näherung an eine Normalverteilung vor. Man beachte, dass der Fall $N = 1$ keine Schwierigkeiten macht, weil man dann auch höchstens $n = 1$ haben, nämlich eine einzige Kugel herausziehen kann, und es liegt ein besonders triviales Bernoulli-Experiment vor mit $p = 0$ oder $p = 1$, je nach dem, ob $K = 0$ oder $K = 1$.

Für diejenigen Leser, welche damit nicht zufrieden sind, folgt hier noch der Beweis beider Formeln für die hypergeometrischen Verteilungen:

Zu μ: Wir fassen das Experiment zur $(N, K, n)-$ hypergeometrisch verteilten Trefferzahlvariablen X so auf: n mal wird je eine Kugel aus der Urne ohne Zurücklegen gezogen. Wir definieren Variablen X_i so, dass X_i den Wert 1 erhält, wenn die $i-$ te gezogene Kugel eine Trefferkugel ist, sonst erhält X_i den Wert Null. Wir haben dann $X = \sum_{i=1}^{n} X_i$. Folglich ist $\mu(X) = \sum_{i=1}^{n} \mu(X_i)$. Wir zeigen nunmehr, dass $\mu(X_i) = K/N$, für alle i, $1 \leq i \leq n$. Damit folgt dann sofort die Behauptung. Offensichtlich gilt $\mu(X_1) = K/N$, da wir nur eine Kugel aus der Urne ziehen und sich X_1 als $(K/N)-$ Bernoulligröße verhält. Nun setzen wir voraus, dass bereits für $k < n$ gelte: $\mu(X_i) = K/N$ für alle $i \leq k$. Wir haben unter dieser Voraussetzung (vollständige Induktion) zu zeigen, dass auch $\mu(X_{k+1}) = K/N$ gilt - dann ist $\mu(X_i) = K/N$ für alle $i \leq n$ bewiesen. Nach der Induktionsvoraussetzung sind nach k Zügen im Mittel $\mu(\sum_{i=1}^{k} X_i) = k \cdot K/N$ Trefferkugeln gezogen. Im Mittel sind also noch $K - k \cdot K/N = K(N-k)/N$ Trefferkugeln vorhanden, unter noch insgesamt $N - k$ Kugeln. Die Wahrscheinlichkeit, eine Trefferkugel im $(k+1)$. Zug zu erhalten, ist also

$$\frac{K(N-k)}{N(N-k)} = \frac{K}{N}.$$

Also ist dies der Erwartungswert $\mu(X_{k+1})$.

Zu σ^2: Wir machen uns zunächst klar - X und X_i bedeuten dasselbe wie oben, dass

$$(*)\ \sigma^2(X) = \sigma^2\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma^2(X_i) + \sum_{1 \leq i,j \leq n, i \neq j}^{n} Cov(X_i, X_j),$$

in Verallgemeinerung der Formel $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + 2Cov(X,Y)$. Nun stellen wir aber fest, dass für alle $i \neq j, 1 \leq i, j \leq n$ der Wert von $Cov(X_i, X_j)$ stets derselbe ist, und zwar

$$Cov(X_i, X_j) = \frac{K}{N} \cdot \frac{K-N}{N(N-1)}.$$

Denn (sei $i \neq j$) $X_i X_j$ hat genau dann den Wert 1, wenn beide Kugeln Trefferkugeln sind. Die Wahrscheinlichkeit dafür, dass die $i-$ te eine solche ist, beträgt K/N, wie oben eingesehen, und wenn die $i-$ te eine Trefferkugel ist, dann hat man eine bedingte Wahrscheinlichkeit von $(K-1)/(N-1)$ dafür, dass auch X_j den Wert 1 erhält. Also nimmt $X_i X_j$ den Wert 1 genau mit der Wahrscheinlichkeit

$$\frac{K}{N} \cdot \frac{K-1}{N-1}$$

an, und somit ist

$$\begin{aligned} Cov(X_iX_k) &= \mu(X_iX_k) - \mu(X_i)\mu(X_k) \\ &= \frac{K}{N}\cdot\frac{K-1}{N-1} - \frac{K}{N}\cdot\frac{K}{N} \\ &= \frac{K}{N}\cdot\frac{K-N}{N(N-1)}. \end{aligned}$$

Wie zu erwarten, ist sie im allgemeinen negativ! Weiter stellt man fest dass

$$\sigma^2(X_i) = \frac{K}{N}\cdot\frac{N-K}{N-K} \text{ für } 1 \le i \le n,$$

man denke stets an die einfache Bernoulli-Situation mit $p = K/N$.

Insgesamt erhalten wir damit aus (*), da es genau $n(n-1)$ Zahlenpaare (i,j) mit $1 \le i,j \le n$ und $i \neq j$ gibt, dass

$$\begin{aligned} \sigma^2(X) &= n\frac{K}{N}\left(1-\frac{K}{N}\right) + n(n-1)\frac{K}{N}\cdot\frac{K-N}{N(N-1)} \\ &= n\frac{K}{N}\left(\frac{N-K}{N} - (n-1)\frac{N-K}{N(N-1)}\right) \\ &= n\frac{K}{N}\cdot\frac{N-K}{N}\left(1-\frac{n-1}{N-1}\right) \\ &= n\frac{K}{N}\cdot\frac{N-K}{N}\left(\frac{N-1-(n-1)}{N-1}\right) \\ &= n\frac{K}{N}\cdot\frac{N-K}{N}\cdot\frac{N-n}{N-1}. \end{aligned}$$

KAPITEL 3

Vertrauensintervalle und Hypothesentests

Bisher haben wir in gewissen Situationen untersucht, welche Wahrscheinlichkeiten für interessierende Ereignisse bestehen. Es wurde dabei stets vorausgesetzt, dass jeweils die Modellbeschreibung zutraf, somit auch die auf dieser Beschreibung fußende mathematische Wahrscheinlichkeitsberechnung. Außerdem haben wir empirische Verteilungen im Sinne deskriptiver Statistik einfach beschrieben. Beide Enden fassen wir nunmehr zu einer neuen Sichtweise zusammen: Stellen wir uns vor, dass eine empirische Situation zu beschreiben sei, dass wir aber weder ein gültiges mathematisches Modell dafür haben, aus dem einfach zu folgern wäre, noch eine vollständige empirische Datenerfassung, die dann einfach zu komprimieren wäre. Natürlich benötigen wir wenigstens *einige* Daten - sonst werden wir schwerlich begründete Aussagen machen können, aber wir nehmen nunmehr nur an, im Besitze einer Stichprobe aus der Population zu sein, dass für diese Stichprobe die Werte einer Variablen X (oder auch mehrerer Variablen $X, Y, ...$) uns bekannt seien. Es liegt naiv nahe, die Verhältnisse innerhalb der Stichprobe auf die Gesamtpopulation zu übertragen, aber da liegt das Problem: Inwieweit ist diese Übertragung erlaubt, d.h. genau und sicher genug? Das werden im Folgenden die beiden Momente sein, nach denen stets zu fragen ist: Sicherheit und Genauigkeit. Ein Beispiel: Wenn wir von 1000 zufällig ausgewählten Bundesrepublikanern wissen, ob sie Blutgruppe A oder eine andere haben, so wird mit hoher Sicherheit, d.h. Wahrscheinlichkeit der in dieser Stichprobe angetroffene Anteil von Blutgruppe A in einem gewissen (nicht allzu weiten) Abstand vom betreffenden Anteil in der Gesamtbevölkerung sein. Sicherheit und zugleich Genauigkeit werden besser sein können als bei einer Stichprobe von nur 100 Bürgern. Aber wie sicher und wie genau? Gerade dies wollen wir *quantitativ* bestimmen. Das ist zunächst einmal der Beitrag der „Schließenden Statistik" oder Inferenzstatistik, die eben von Stichproben mit gewisser angebbarer Sicherheit und Genauigkeit auf die Population schließt.

Zunächst einmal wollen wir klarstellen, dass man im allgemeinen nur völlig unbrauchbare, weil viel zu ungenaue Aussagen machen kann, wenn man *absolute* Sicherheit verlangt: Haben wir etwa die Daten von 1000 Bundesbürgern und dabei 30% mit Blutgruppe A angetroffen, so wissen wir sicher nur, dass mindestens 300/80000000, also knapp 4 Millionstel der Bundesbürger diese Blutgruppe hat. Verzichten wir dagegen auf absolute Sicherheit und begnügen uns mit einer hohen Wahrscheinlichkeit unserer Aussage, so können wir dagegen eine viel brauchbarere, weil genauere Aussage machen. Im vorliegenden Fall z.B. bekommt man mit den Mitteln dieses Kapitels heraus, dass mit einer Sicherheit von 0.99 oder 99% gilt: Der Anteil der Leute mit Blutgruppe A liegt im Bereich $[0.26; 0.34]$, also zwischen 26 und 34 Prozent. Verlangen wir eine höhere Sicherheit, also etwa 0.999 oder 99.9%, so können wir nur eine ungenauere Aussage treffen, ber immerhin noch sagen: Zwischen 25% und 35%. Das mag wie Zauberei anmuten, da wir doch nur

einen verschwindenden Teil der Population mit unserer Stichprobe untersucht haben, es ist aber reine Wahrscheinlichkeitsrechnung, und die erweist, dass in die Berechnung der Wahrscheinlichkeit oder Sicherheit zu vorgegebener Genauigkeit (bzw. umgekehrt) keineswegs das Verhältnis des Stichprobenumfangs zum Populationsumfang eingeht, sondern allein der absolute Stichprobenumfang entscheidet. Allerdings muss hier angemerkt werden, dass die Gewinnung einer „sauberen" Zufallsstichprobe aus einer Riesenpopulation nicht ganz einfach praktisch durchzuführen ist. Über diesen Weg schlägt das Problem denn doch noch zu, es kann aber immerhin wirksam angegangen werden im Gegensatz zu einer Generalerfassung der Bevölkerung.

Im Beispiel haben wir zu verschiedenen Sicherheiten zwei Vertrauensintervalle angegeben für die Lage eines unbekannten Populationsmittelwertes (nämlich der Bernoulligröße, welche jedem Populationsmitglied den Wert 1 zuordnet, falls es Blutgruppe A hat, sonst den Wert 0 - der Mittelwert μ dieser Größe ist $p =$ relative Häufigkeit der Blutgruppe A in der Bevölkerung). Ähnlich gelagert sind Probleme der Form, dass eine Hypothese H_0 über eine Variable X in einer Population Ω empirisch zu testen ist anhand einer Zufallsstichprobe von $X-$ Werten. - Eine Hypothese ist zunächst einmal eine Aussage, und zwar eine solche, deren Wahrheit fraglich ist (aber möglichst auch interessant!). (Es kann sich natürlich auch um eine Hypothese über eine Beziehung zwischen mehreren Variablen handeln, dann müssten all diese in der Stichprobe erhoben sein.) Man kann dann H_0 nicht direkt und sicher überprüfen - dazu müsste man den überwältigenden Teil der Population kennen, aber man kann aus H_0 folgern, dass ein Ereignis B in der Stichprobe mit hoher Wahrscheinlichkeit zu beobachten sein müsste, und dann die Hypothese H_0 als ihrerseits sehr unwahrscheinlich verwerfen, wenn man in der Stichprobe $\overline{B}$ beobachtet hat. Geben wir auch dazu ein Beispiel, diesmal eines, das wir bereits mit den bisherigen Mitteln rechnen können: Sei H_0 die Hypothese, unter den Wuppertaler Studenten gebe es höchstens 5%, die ihr Studium ausschließlich mit eigener Werktätigkeit finanzieren. In einer Zufallsstichprobe von 10 Studenten fanden Sie aber 3, welche dies tun. In der Stichprobe liegt der Anteil also deutlich höher (0.3), als die Hypothese für die Population sagt (0.05). Können wir das auf die Population übertragen, also die Hypothese verwerfen? Es fragt sich hier - wie immer in diesem Kontext, ob die Abweichung zufällig so hoch sein könnte oder ob man eher annehmen sollte, dass andere Verhältnisse in der Population vorliegen. Eine Abweichung gibt es praktisch immer - in unserem Falle wären genau 5% in der Stichprobe niemals zu beobachten! Stellen wir uns nun auf den intuitiven Standpunkt, dass uns die Beobachtung von 3 Selbstfinanzierern unter 10 Leuten reicht, die Hypothese zu verwerfen. Sicher hätten wir dann auch bei 4,5,...,10 beobachteten Selbstfinanzierern verworfen. Das ist nun das Ereignis $\overline{B}$: Trefferzahl mindestens 3 (unter 10 zufällig Ausgewählten). Welche Wahrscheinlichkeit lässt sich aus der Hypothese H_0 für B folgern? Das ist gemäß Binomialverteilung $\sum_{k=0}^{2} \binom{10}{k} 0.05^k \cdot 0.95^{10-k} = 0.9885$. Setzen wir die Hypothese als wahr voraus, so erhalten wir also eine Wahrscheinlichkeit von fast 99% für B, also von kaum mehr als einem Prozent für $\overline{B}$, man würde also sagen, dass die Beobachtung gegen die Hypothese spreche. Natürlich sind wir nicht absolut *sicher*, dass die Hypothese falsch sein müsse - sie schließt das Beobachtete nicht aus, aber sollte uns gerade so Unwahrscheinliches passiert sein? Man beschreibt die Entscheidung gegen die Hypothese in solchem Falle so: „Die Hypothese wird auf dem Signifikanzniveau 1% (oder Niveau 0.01) verworfen". Das

bedeutet: Eine mindestens so große Abweichung zwischen Beobachtung und Hypothese wie tatsächlich beobachtet hat laut Hypothese eine Wahrscheinlichkeit unter 1%. (Im Beispiel reichte es nicht ganz dafür, sondern nur zum Niveau 0.012.) Das Signifikanzniveau gibt also an, mit welcher Wahrscheinlichkeit folgender sogenannter Fehler erster Art begangen wird: Die Hypothese wird verworfen, ist aber wahr. Genauer handelt es sich um die Wahrscheinlichkeit dafür, dass ein zum Verwerfen führendes Beobachtungsergebnis herauskommt, *wenn die Hypothese wahr ist,* also unter der Bedingung der Hypothese. Wichtig ist die Spezifikation des Ereignisses B vorab; denn sonst könnte man einfach irgendeine Besonderheit des beobachteten Resultats auswählen, die laut Hypothese sehr unwahrscheinlich wäre, aber eben nicht laut wegen des Inhalts der Hypothese, sondern weil sie stets sehr unwahrscheinlich ist. Beispiel: Es soll ein Würfel getestet werden daraufhin, ob er ein Sechstel Wahrscheinlichkeit für Sechsen hat. Unter 600 Würfen beobachtet man genau 100 Sechsen. Das ist aber sehr unwahrscheinlich (obgleich genau der Erwartungswert laut Hypothese), Wahrscheinlichkeit $\binom{600}{100}\left(\frac{1}{6}\right)^{100}\left(\frac{5}{6}\right)^{500} \approx 0.044$. Merke: Das Konkrete ist immer beliebig unwahrscheinlich, es spricht nicht gegen eine allgemeine Hypothese. Im Beispiel würde man allerdings sehen, dass alle denkbaren Alternativen zur Hypothese das Beobachtete noch unwahrscheinlicher machen würden.

Man kann zu allen erdenklichen Sachverhalten Hypothesen aufstellen - natürlich sollte man das vernünftigerweise erst nach gewisser Erfahrung und Kenntnis tun. Nicht alle davon sind statistisch zu prüfen, es gibt auch andere, z.B. die Hypothese, die Erde sei eine Kugel: Diese wurde vor beinahe 2000 Jahren auf raffinierte Weise dadurch geprüft, dass man die Einfallswinkel der Sonnenenstrahlen an zwei 1000 km entfernten Orten maß - man konnte sogar den Erdradius erstaunlich genau damit angeben. Statistisch zu prüfen sind zunächst einmal nur solche Hypothesen, die sich auf Zufallsvariablen beziehen. Aber auch von solchen Hypothesen gibt es eine große Vielfalt, aus der wir nur einige der wichtigsten Beispiele bringen. Ebenso vielfältig ist das Bestimmen von Vertrauensintervallen für unbekannte Parameterwerte von Verteilungen. Wir gehen nun so vor: Zunächst werden diese beiden Aufgabenstellungen abstrakt beschrieben, anschließend werden sie durchgeführt für den einfachsten Fall, in dem es um einen unbekannten Populationsmittelwert $\mu(X)$ geht sowie um den Vergleich zweier Populationsmittelwerte $\mu(X_{|\Omega_1})$ und $\mu(X_{|\Omega_2})$.

1. Abstrakte Beschreibung der Aufgaben

1.1. Das Schätzen eines unbekannten Verteilungsparameters. Man hat eine Variable X und möchte einen Parameter, nennen wir ihn allgemein *par*, der Verteilung von X anhand einer Stichprobe schätzen, etwa $\mu(X)$ oder $\sigma(X)$ oder den Median von X. Dann sucht man sich eine Schätzvariable $\mathcal{S}$, von der ein Wert anhand der Stichprobe berechnet und damit indirekt „beobachtet" werden kann. Mittels der mathematischen Theorie zum wahrscheinlichkeitstheoretischen Verhalten von $\mathcal{S}$ bestimmt man dann ein Vertrauensintervall zu $\mathcal{S}$ der Art: „Der Wert von $\mathcal{S}$ liegt mit Sicherheit w im Bereich $par \pm \varepsilon$". Dann weiß man auch, dass der unbekannte Parameter *par* vom beobachteten Wert $\mathfrak{s}$ von $\mathcal{S}$ nur höchstens den Abstand ε hat, mit eben der Sicherheit w, die eine Wahrscheinlichkeit nahe 1 sein sollte.

1.2. Statistisches Testen einer Hypothese H_0 anhand einer Stichprobe.

1. Man formuliert sorgfältig die zu prüfende Hypothese H_0.

2. Man bestimmt ein Signifikanzniveau α, das ist eine vorgeschriebene kleine Wahrscheinlichkeit, mit der beim nunmehr zu besprechenden Verfahren herauskommt, dass die Hypothese *fälschlich verworfen,* also der Fehler 1. Art begangen wird: Diese Wahrscheinlichkeit ist unter Voraussetzung der Hypothese zu verstehen.
3. Man bestimmt *unter Voraussetzung der Hypothese* ein Vertrauensintervall für eine sogenannte Prüfgröße T, d.h. einen Bereich, in dem ein zufällig beobachteter Wert von T mit der Wahrscheinlichkeit $1-\alpha$ liegt. (Das Ereignis B aus der Einleitung ist dann gerade „$a \leq T \leq b$" oder auch „$a \leq T$", „$T \leq b$". Der Bereich ist also im allgemeinen ein Intervall. Dies Ereignis sagt die Hypothese mit hoher Wahrscheinlichkeit voraus. Die Form des Bereiches richtet sich nach der Form der Hypothese H_0, s.u. unter „Einseitige und zweiseitige Hypothesenformulierung".)
4. Man beobachtet einen Wert t der Testvariablen T (gewöhnlich anhand einer Stichprobe).
5. Entscheidung: Liegt t nicht im Bereich aus 2., also nicht im Vertrauensintervall, sondern im Rest, dem sog. „Verwerfungsbereich", so ist also $\overline{B}$ eingetreten, und die Hypothese H_0 wird auf dem Niveau α verworfen. (Das bedeutet: Man ist recht sicher, dass H_0 falsch sein muss, folglich die Verneinung H_1 von H_0 wahr. Man sagt daher auch *mit Recht*, H_1 werde *angenommen* - eben als wahr angenommen. Liegt t dagegen im Vertrauensintervall, so sage man präzise: Die Beobachtung reicht nicht aus, um die Hypothese H_0 mit hoher Sicherheit $(1-\alpha)$ (bzw. geringer Irrtumswahrscheinlichkeit α) zu verwerfen. *Es ist eine leider immer wieder begegnende komplette Fehlinterpretation,* dies zu verwechseln damit, H_0 sei nunmehr als wahr „anzunehmen", oder gar, H_0 sei daher mit Sicherheit $1-\alpha$ wahr.

Wir werden das Schema ausführlich vor allem im Beispiel der Hypothesen über Populationsmittelwerte ausfüllen. Hier begnügen wir uns damit, ausdrücklich zu begründen, warum die zuletzt genannte Fehlinterpretation wirklich kompletter Unsinn ist: Stellen Sie sich vor, H_0 sei die Hypothese, die mittlere tägliche Fernsehzeit bei Jugendlichen liege bei 2 Stunden. Sie haben in einer Stichprobe ein Mittel von 2.5 Stunden beobachtet, aber das Vertrauensintervall zu 0.99 reiche laut Hypothese bis 2.6 Stunden. Sie können also nicht auf Niveau 0.01 verwerfen. Aber auf dem Niveau 0.05 könnten Sie verwerfen, das Vertrauensintervall dafür reiche nur bis 2.4. Dann hätten Sie immer noch die Sicherheit von 0.95 dafür, dass die Hypothese *falsch sei* und *nicht etwa* eine hohe Sicherheit dafür (gar 0.99), dass die Hypothese *richtig* sei. Genau dies ist aber der Inhalt jener Fehlinterpretation. Auch in der schwächeren Form, man „nehme die Hypothese H_0 an", bleibt sie Unsinn. Man sollte einen intuitiven Begriff von „schwachen" und „starken" Tests haben: Bei kleinem Stichprobenumfang liegt ein schwacher Test vor - fällt eine Hypothese bereits bei einem solchen um, so ist sie ziemlich sicher falsch. Fällt sie aber dabei nicht, so hat sie erst einen sehr schwachen Test bestanden - ein stärkerer könnte sie durchaus noch zu Fall bringen, es gibt noch lange keinen Grund dafür, sie „anzunehmen" oder zu glauben. Besteht eine Hypothese dagegen einen starken Test, so kann man sehr wohl glauben, dass die Hypothese zumindest nicht allzu falsch sein kann.

Das ergibt nur scheinbar ein Dilemma: Häufig werden wir eine Hypothese H im Auge haben, von der wir die Wahrheit zeigen wollen. Bei statistischem Test

von $H_0 = H$ kann man nach dem Vorangehenden nicht so verfahren, dass man nur H_0 nicht verwirft. Stattdessen kann man die Sache oft so angehen: Man bildet die Verneinung „nicht H" und nennt *diese* H_0, steckt sie in einen statistischen Test. Kommt man zum Verwerfen von H_0 auf vernünftigem Niveau, so ist H_0 also ziemlich sicher falsch und H_1 ziemlich sicher wahr. Das Verfahren klappt nicht immer (s.u.), aber im ungünstigen Fall kann man durch einen starken Test unter Kontrolle des Fehlers 2. Art (Hypothese ist falsch - mit spezifizierter Abweichung von der Wahrheit, wird aber fälschlich nicht verworfen) doch zu einem quantifizierten Bestätigungsresultat kommen. Allerdings wird man vielfach einfacher vorgehen können und etwa ein Vertrauensintervall angeben für den fraglichen Verteilungsparameter (ein solcher ist vielfach der Gegenstand einer Hypothese). Hat man etwa für die mittlere Fernsehzeit der Jugendlichen ein sehr sicheres Vertrauensintervall von 2 bis 2.3 Stunden bestimmt, so kann eine Hypothese H des Inhalts, dieser Mittelwert liege bei 2.25, nicht allzu falsch sein, also im Rahmen einer allenfalls anzustrebenden Genauigkeit richtig. Auch eine Angabe von 2.5 wäre noch korrekt, wenn Differenzen unter 0.5 Stunden nicht interessieren.

2. Konkretisierung für den Fall eines unbekannten $\mu(X)$

2.1. Bestimmung eines Vertrauensintervalls für einen unbekannten Populationsmittelwert $\mu(X)$. Die naheliegend zu verwendende Schätzgröße $\mathcal{S}$ ist in diesem Fall $\overline{X}$. Der Grund ist folgender: Man hat $\mu(\overline{X}) = \mu(X)$, und $\sigma(\overline{X}) = \sigma(X)/\sqrt{n}$,wie im vorigen Abschnitt eingesehen. Die Werte von $\overline{X}$ tendieren also zu $\mu(X)$, desto stärker, je größer der Stichprobenumfang n wird. Mit einer Stichprobe $x_1, ..., x_n$ von $X-$ Werten können wir davon den Wert beobachten:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Ebenso wichtig: Die Variable $\overline{X}$ ist selbst bei bescheidenen Stichprobenumfängen bereits näherungsweise normalverteilt, wie aus dem Zentralen Grenzwertsatz folgt, so dass man also die zugehörige Wahrscheinlichkeitsrechnung gut beherrscht, sobald man μ, σ kennt. Daher lautet das zweiseitige Vertrauensintervall zur Sicherheit w für $\mu(X)$:

$$\overline{x} \pm z_{w+(1-w)/2} \cdot \frac{\sigma(X)}{\sqrt{n}} \tag{2.1}$$

Dabei ist z_w zu einer Wahrscheinlichkeit w die Grenze z, für die $\Phi_{0,1}(z) = w$ gilt, also die in der Tabelle zu w abzulesende Grenze. Die Begründung zur Formel: Zunächst haben wir für die Variable $\overline{X}$ das Vertrauensintervall

$$\mu(X) \pm z_{w+(1-w)/2} \cdot \frac{\sigma(X)}{\sqrt{n}}. \tag{2.2}$$

Denn

$$\begin{aligned}
&\Phi_{\mu(X),\frac{\sigma(X)}{\sqrt{n}}}\left(\mu(X)+z_{w+(1-w)/2}\cdot\frac{\sigma(X)}{\sqrt{n}}\right)\\
&-\Phi_{\mu(X),\frac{\sigma(X)}{\sqrt{n}}}\left(\mu(X)-z_{w+(1-w)/2}\cdot\frac{\sigma(X)}{\sqrt{n}}\right)\\
=\;& 1-2\Phi_{\mu(X),\frac{\sigma(X)}{\sqrt{n}}}\left(\mu(X)-z_{w+(1-w)/2}\cdot\frac{\sigma(X)}{\sqrt{n}}\right)\\
=\;& 1-2\Phi_{0,1}(z_{w+(1-w)/2})=w.
\end{aligned}$$

Nunmehr die Überlegung: Formel 2.2 bedeutet, dass ein beliebig „gezogener" Wert von $\overline{X}$ mit Wahrscheinlichkeit w um höchstens den Betrag $z_{w+(1-w)/2}\cdot\frac{\sigma(X)}{\sqrt{n}}$ von $\mu(\overline{X})=\mu(X)$ abweicht. Also weicht auch $\mu(X)$ um höchstens diesen Betrag von dem beobachteten Wert $\overline{x}$ ab, mit der Sicherheit w. Daher Formel 2.1.

Nun hat die gewonnene Lösung des Problems einen Schönheitsfehler: Sie verlangt Eingabe des Wertes $\sigma(X)$, der natürlich ebenso unbekannt ist wie $\mu(X)$. Diese Schwierigkeit kann man auf zwei Weisen überwinden:

- 1. Möglichkeit: Man ersetzt $\sigma(X)$ durch eine obere Schranke $m\geq\sigma(X)$, von der man theoretisch/empirisch weiß, dass $\sigma(X)$ darunter bleiben muss. Allerdings hat man dann ein Vertrauensintervall zu irgendeiner Wahrscheinlichkeit $w'\geq w$ bestimmt. Allerdings hat man sich, um nur Wahrscheinlichkeit w zu behaupten, eventuell eine zu große Ungenauigkeit eingehandelt - die halbe Breite des Vertrauensintervalls heißt nun $z_{w+(1-w)/2}\cdot\frac{m}{\sqrt{n}}$ und ist vielleicht unnötig groß. Ein Beispiel: Sei X eine $p-$ Bernoulli-Größe, mit unbekanntem $\mu(X)=p$. Dann ist $\overline{X}$ die Größe: Relative Häufigkeit der Treffer in der Stichprobe. Nun gilt $\sigma(X)=\sqrt{p(1-p)}$, und das ist stets höchstens $\sqrt{\frac{1}{2}\cdot\frac{1}{2}}=\frac{1}{2}$. In diesem Falle kann man daher $m=1/2$ setzen. Hat man also z.B. in einer Stichprobe vom Umfang 100 genau 30 Treffer gefunden, so erhält man mit $0.3\pm 2.58\frac{1}{2\sqrt{100}}$, also etwa $[0.17;0.47]$ als 99%-Vertrauensintervall.
- 2. Möglichkeit: Man ersetzt den unbekannten Wert $\sigma(X)$ durch den geeigneten Schätzwert

$$s(X):=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}. \tag{2.3}$$

Entsprechend lautet der Schätzwert für die Streuung $\sigma(\overline{X})$, mit der wir letztlich arbeiten:

$$s(\overline{X}):=\frac{s(X)}{\sqrt{n}}. \tag{2.4}$$

Man beachte: Das geht wie die Berechnung der Streuung innerhalb der Stichprobe, diese als Population aufgefasst - anders nur: Der Faktor $1/(n-1)$ anstelle von $1/n$. (Man kann zeigen, dass mit $1/n$ die Streuung systematisch unterschätzt würde.) In diesem Falle ersetzt man also $\sigma(X)$ durch den Wert einer neuen Zufallsgröße $S(X)$, und dieser Wert kann zufällig eben größer oder kleiner als $\sigma(X)$ sein. Aus diesem Grunde ergibt $\overline{x}\pm z_{w+(1-w)/2}s(X)/\sqrt{n}$ kein exaktes Vertrauensintervall mehr zur Wahrscheinlichkeit w, und die Abweichung hängt natürlich von der Sicherheit des Schätzers $S(X)$ ab, die mit

steigendem Stichprobenumfang besser wird. Aber es gibt eine einparametrige Schar von Verteilungen, in der gerade der Parameter n vorkommt und die gerade die Kompensation für das Einsetzen von $s(X)$ für $\sigma(X)$ leistet: Das sind die (wiederum symmetrisch um den Mittelwert 0 liegenden!) $t-$ Verteilungen mit jeweils $1, 2, ...$ Freiheitsgraden, wobei gilt:

$$\textit{Anzahl der Freiheitsgrade} = n - 1.$$

Nun läuft das Verfahren sehr einfach: Man hat mit der $t-$ Tabelle anstelle der $z-$ Tabelle zu arbeiten, also das Resultat:

Das zweiseitige Vertrauensintervall zur Wahrscheinlichkeit w
für unbekannten Mittelwert $\mu(X)$ lautet
$\overline{x} \pm t^{n-1}_{w+(1-w)/2} s(\overline{X})$ (Index $n-1$ für die Freiheitsgrade, auch „df" genannt).

Man wird in einer groben Tabelle ab $n-1 = 100$ bzw. 200 keinen Unterschied zur Standard-Normalverteilung mehr bemerken, dagegen große Unterschiede bei kleinen Stichprobenumfängen n. Für dasselbe Beispiel wir oben hätte man das Vertrauensintervall $0.3 \pm t^{99}_{0.995}\sqrt{\frac{0.3 \cdot 0.7}{99}}$, das ist $0.3 \pm 2.63\sqrt{\frac{0.3 \cdot 0.7}{99}}$, also ungefähr $[0.18, 0.42]$. Man sieht: Der benötigte $t-$ Wert ist mit 2.63 etwas höher als der $z-$Wert 2.58. Aber der Schätzwert $s(X)$ wird kleiner als die obere Schranke $1/2$, und wir erhalten ein etwas günstigeres Vertrauensintervall als oben. Wir haben dabei folgendes nützliche Resultat benutzt, das stets für Bernoulligrößen X gilt:

Für Bernoulligrößen X gilt stets: $s(\overline{X}) = \dfrac{s(X)}{\sqrt{n}} = \sqrt{\dfrac{\overline{x}(1-\overline{x})}{n-1}}$.

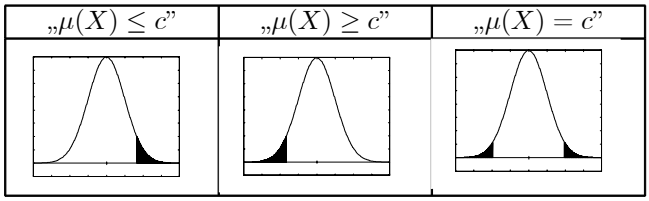
Dabei ist $\overline{x}$ die in der Stichprobe beobachtete relative Häufigkeit der Treffer. (In andern Fällen muss man $s(X)$ aufgrund der beobachteten Einzelwerte ausrechnen.)

2.2. Test einer Hypothese über einen unbekannten Populationsmittelwert $\mu(X)$.

- Formulierungen von H_0:

Einseitig	Zweiseitig
„$\mu(X) \leq c$" „$\mu(X) \geq c$"	„$\mu(X) = c$"

„c" bedeutet dabei einfach den beliebigen Grenzwert bzw. Wert, der in der jeweiligen Hypothese behauptet wird.
- Vorgabe des Signifikanzniveaus α
- Zugehörige Formen von Vertrauensintervallen zu $1 - \alpha$ und entsprechend Verwerfungsbereichen zu α - die schwarz markierten Teile zeigen die jeweils zum Verwerfungsbereich gehörigen Wahrscheinlichkeiten - stets c in der Mitte:



- Berechnung der Vertrauensintervallgrenzen:

Hypothese	„$\mu(X) \leq c$"	„$\mu(X) \geq c$"	„$\mu(X) = c$"
Vertr.-Int. bei bek. $\sigma(\overline{X})$	$]-\infty, c + z_{1-\alpha} \cdot \sigma(\overline{X})]$	$[c + z_{\alpha} \cdot \sigma(\overline{X}), \infty[$	$c \pm z_{1-\alpha/2} \cdot \sigma(\overline{X})$
gewöhnlich: Vert.-Int. mit $s(\overline{X})$ u. $t-$ Vert.	$]-\infty, c + t^{n-1}_{1-\alpha} \cdot s(\overline{X})]$	$[c + t^{n-1}_{\alpha} \cdot s(\overline{X}), \infty[$	$c \pm t^{n-1}_{1-\alpha/2} \cdot s(\overline{X})$

Man beachte, dass $z_\alpha = -z_{1-\alpha}$ und $t^{n-1}_\alpha = -t^{n-1}_{1-\alpha}$. Ferner sehe man ein, dass $\Phi_{0,1}(z_{1-\alpha/2}) - \Phi_{0,1}(-z_{1-\alpha/2}) = 1 - \alpha$ für $1 - \alpha/2 \geq 0.5$ (analog für $t-$ Verteilung), und α sollte im Kontext eine kleine Wahrscheinlichkeit sein. Erinnerung: $\sigma(\overline{X}) = \sigma(X)/\sqrt{n}$ mit n = Stichprobenumfang. Man stößt auch hier im allgemeinen auf das bereits oben behandelte Problem des unbekannten Wertes von $\sigma(X)$, und man löst es wie oben, arbeitet also bei Einsetzen von $s(\overline{X}) = s(X)/\sqrt{n}$ für $\sigma(\overline{X})$ mit $t^{n-1}_{\dots}$anstelle von $z_{\dots}$. Das steht in der dritten Zeile. Speziell z.B. im Falle einer Bernoulligröße X hat man jedoch den günstigen Umstand, dass die Hypothese über $\mu(X)$ bereits den Wert von $\sigma(X)$ impliziert - dann ist der zu nutzen, und man kann gemäß der zweiten Zeile des Schemas verfahren.

- Entscheidung: Verwerfen auf Niveau α, wenn $\overline{x}$ nicht im Vertrauensbereich, also im Verwerfungsbereich liegt. Wenn dagegen $\overline{x}$ *im Vertrauensintervall* liegt, so sage man *nicht* (s.o.), die Hypothese H_0 werde „angenommen" - darum wird auch das Vertrauensintervall laut Hypothese hier *nicht „Annahmebereich"* genannt.

2.2.1. *Eine Modifikation des Schemas.* Zuweilen zieht man es vor, kein Standard-Signifikanzniveau wie $\alpha = 0.05, 0.01, 0.001$ vorzugeben, sondern umgekehrt danach zu fragen, auf welchem Signifikanzniveau man die Hypothese anhand des Beobachteten gerade noch verwerfen könnte. Man berechnet also nicht die Vertrauensgrenze zu α, sondern den Wert α, welcher zur Beobachtung $\overline{x}$ als Vertrauensgrenze gehört. Dann gibt man das Resultat in der Form „$\alpha = ...$", „$\alpha < ...$" oder „$p = ...$" an (wobei *in diesem Kontext* „p" dasselbe wie „α" bedeutet) - so z.B. in vielen Publikationen und auch Computerausdrucken von Statistikprogrammen.

2.2.2. *Beispiele.*

1. X sei die Variable „Intelligenzquotient" (im Sinne einer bestimmten Messung) in einer speziellen Population von Kindern, H_0 sei die Hypothese „$\mu(X) \geq 110$", beobachtet habe man in einer Stichprobe vom Umfang 100 aus jener Population $\overline{x} = 106$ und $s(X) = 10$.
   (a) Erste Version: Test von H_0 auf vorgegebenem Niveau $\alpha = 0.01$. Man errechnet mittels der benötigten Zahl $t^{99}_{0.01} = -2.365$ die Vertrauensgrenze zu.—

$$110 - 2.365\frac{10}{\sqrt{100}} = 107.64$$

   und verwirft also H_0 auf dem Niveau $\alpha = 0.01$. Entsprechend ist man recht sicher, dass gilt: $\mu(X) < 110$.

(b) Zweite Version: Angabe des besten Niveaus, auf dem man H_0 mittels der Beobachtung gerade noch verwerfen kann. Man löst die Gleichung

$$110 + t\frac{10}{\sqrt{100}} = 106$$

mit dem Resultat $t = -4$ und fragt nunmehr nach der Wahrscheinlichkeit, die links von diesem Wert sitzt. Zu $t^{99} = -4$ gehört eine Wahrscheinlichkeit $\alpha < 0.0000612$, man kann also auf dem (sehr guten) Niveau 0.0000612 die Hypothese H_0 verwerfen und ist in der Tat *sehr sicher*, dass $\mu(X) < 110$. Dies hätte man bei der ersten Version nicht so genau gesehen, sondern nur grob geahnt, dass man das Niveau 0.01 bei weitem übertreffen könnte.

2. H_0 sei die Hypothese, dass in einer gewissen Population eine Eigenschaft mit relativer Häufigkeit 0.3 anzutreffen sei, also H_0 : „$\mu(X) = 0.3$". Hier ist X die Bernoulligröße, welche jedem Populationsmitglied mit der besagten Eigenschaft die Zahl Eins zuordnet, jedem sonstigen die Zahl Null. Man habe in einer Stichprobe vom Umfang 50 genau 25 mit jener Eigenschaft angetroffen. Das bedeutet: $\overline{x} = 0.5$. Man beachte: Die Hypothese impliziert, dass $\sigma(X) = \sqrt{0.3 \cdot 0.7}$, also $\sigma(\overline{X}) = \sqrt{0.3 \cdot 0.7}/100$. Somit haben wir hier *nicht* mit $t-$ Verteilung zu arbeiten, sondern mit (laut Hypothese) korrekter Streuung, also mit Normalverteilung, da wir für die Berechnung die Hypothese zur Voraussetzung nehmen. Wir besprechen wieder die oben beschriebenen Versionen, woraus in diesem Falle jedoch 4 Versionen werden, da wir einmal mit der Variablen der relativen Stichprobenhäufigkeit, das andere Mal mit der binomialverteilten Variablen der absoluten Stichprobenhäufigkeit arbeiten werden - letzteres Verfahren hat den Vorzug, dass man die Rechnung mit Näherung durch Normalverteilung mittels Stetigkeitskorrektur genauer hinbekommt.

(a) Erste Version: $\alpha = 0.05$. Wir errechnen das zweiseitige Vertrauensintervall

$$0.3 \pm 1.96\frac{\sqrt{0.3 \cdot 0.7}}{\sqrt{50}}, \text{ also } [0.17; 0.43].$$

Der beobachtete Wert 0.5 fällt deutlich heraus, wir verwerfen also auf dem Niveau 0.05.

(b) Zweite Version: Wir bestimmen das Signifikanzniveau, auf dem wir verwerfen können, lösen also die Gleichung

$$0.3 + z\frac{\sqrt{0.3 \cdot 0.7}}{\sqrt{50}} = 0.5$$

und finden $z = 3.0861$. Das gesuchte Niveau ist doppelt so groß wie die Wahrscheinlichkeit, die auf den Bereich oberhalb von z entfällt (bei der Standard-Normalverteilung). Also

$$\alpha = 2\Phi_{0,1}(-3.0861) \approx 0.00203.$$

Man kann also auf einem Niveau nahe 2/1000 verwerfen.

(c) Dritte Version: Wie die erste, nur mit Binomialverteilung: Sei Y die Variable „Trefferzahl" auf der Population aller Stichproben des Umfangs 50. Sie ist laut Hypothese $(50, 0.3)-$ binomialverteilt. Mit

$\alpha = 0.05$ (analog zu a.) ergibt sich das Vertrauensintervall für Y laut Hypothese:

$$15 \pm 1.96\sqrt{50 \cdot 0.3 \cdot 0.7}, \text{ also } [8.6489; 21.351].$$

Mit Stetigkeitskorrekur ist die Obergrenze als 21 anzusetzen (hätte man einen Wert *über 21.5* errechnet, so wäre mit dieser Korrektur 22 anzusetzen gewesen). Der Befund lautet wiederum, dass auf dem Niveau $\alpha = 0.05$ zu verwerfen ist.

(d) Vierte Version: Wie die zweite, wieder jedoch mit Binomialverteilung. Das Niveau α ist die Wahrscheinlichkeit dafür, mindestens 25 oder höchstens 5 Treffer zu bekommen bei einer Binomialverteilung mit $n = 50, p = 0.3$. Näherung durch Normalverteilung ergibt dafür *mit Stetigkeitskorrektur*:

$$\begin{aligned} \alpha &= \Phi_{15;\sqrt{50\cdot 0.3\cdot 0.7}}(5.5) + 1 - \Phi_{15;\sqrt{50\cdot 0.3\cdot 0.7}}(24.5) \\ &= 2\Phi_{15;\sqrt{50\cdot 0.3\cdot 0.7}}(5.5) = 2\Phi_{0,1}\left(\frac{5.5-15}{\sqrt{50\cdot 0.3\cdot 0.7}}\right) \approx 0.0034. \end{aligned}$$

Diese Wahrscheinlichkeit ist zwar auch klein, aber deutlich höher (und korrekter!) als die in der zweiten Version (b.) gegebene. (Wenn man hier ohne Stetigkeitskorrektur mit den Grenzen $5, 25$ arbeiten würde, käme genau dasselbe Resultat wie in b.)

Bemerkung: Man hat bei der Variante, die aus der Beobachtung das erreichbare Niveau berechnet, grundsätzlich die Möglichkeit, die Vertrauensintervallform so zu ändern, dass man t bzw. z als Unbekannte setzt (so in b.) und danach auflöst, oder aber die Wahrscheinlichkeit des Verwerfungsbereichs direkt zu berechnen (so in d.) - beide ergeben dasselbe Resultat.

Bemerkung: Feinere Wertetabellen als die üblichen zur Normalverteilung und $t-$ Verteilung erhält man besonders bequem mit dem Computer - davon wurde in den Beispielen oben freier Gebrauch gemacht.

2.3. Test einer Hypothese über die Differenz zweier unbekannter Populationsmittelwerte. Vielfach interessiert die Frage, ob eine Variable X in einer Teilpopulation A einen höheren / niedrigeren / gleichen Mittelwert wie in einer anderen Teilpopulation B habe. Genauer: Die Variable X sei auf Ω definiert, A und B disjunkte Klassen von Ω, also $A, B \subseteq \Omega$ und $A \cap B = \varnothing$. (Nicht notwendig $A \cup B = \Omega$.) Dann betrachten wir die *verschiedenen Variablen* $X_{|A}$ und $X_{|B}$, das sind die Einschränkungen von X jeweils auf die Definitionsbereiche A, B, also z.B. $X_{|A} : A \to \Omega$, $a \mapsto X(a)$. Man denke etwa an eine medizinisch relevante Variable X und an A, B als Teilpopulationen von Kranken verschiedener Krankheiten usw. Nun möchte man eine Hypothese der Form „$\mu(X_{|A}) \leq [=, \geq] \mu(X_{|B})$" testen, in diesem Sinne einen Mittelwertvergleich anstellen. Der empirische Hintergrund sollte ausgemacht werden von einer Stichprobe eines Umfangs n_A von Messungen in A und einer unabhängige Stichprobe eines (nicht notwenig gleichen) Umfangs n_B aus B. (Die Unabhängigkeit ergibt sich hier normalerweise aus der Grundsituation, dass $A \cap B = \varnothing$. Anders liegt die Sache bei sogenannten „verbundenen" Stichproben, wenn man etwa dieselben Menschen zu verschiedenen Zeitpunkten beobachtet, vor und nach einem Training z.B. Jedenfalls wollen wir so vorsichtig sein, diese Unabhängigkeit der Stichproben ausdrücklich zu fordern.)

Die Lösung des Problems besteht in der Zurückführung auf den Fall einer einzigen Variablen, der oben bereits behandelt wurde. Allerdings ergibt sich noch ein gesondertes kleines technisches Problem mit der Streuungsschätzung.

Die Zurückführung: Man nutzt die Gleichwertigkeit von z.B. „$\mu(X) \leq \mu(Y)$" mit „$\mu(X) - \mu(Y) \leq 0$", dann $\mu(X) - \mu(Y) = \mu(X - Y)$, schließlich $\mu(X) = \mu(\overline{X})$ und erhält - den dritten Fall kann man sich schenken, da man nur A und B zu vertauschen braucht, um erstere Version wieder zu erhalten:

	Hypothese einseitig	Hypothese zweiseitig
Hypothese	„$\mu(X_{\mid A}) \leq \mu(X_{\mid B})$"	„$\mu(X_{\mid A}) = \mu(X_{\mid B})$"
gleichwertige Umformulierung	„$\mu(\overline{X_{\mid A}} - \overline{X_{\mid B}}) \leq 0$"	„$\mu(\overline{X_{\mid A}} - \overline{X_{\mid B}}) = 0$"

Setzen wir nun $Y := \overline{X_{\mid A}} - \overline{X_{\mid B}}$, so haben wir die Vergleichs-Hypothesen auf eine über Y zurückgeführt, welche die Form einer Hypothese über einen einzelnen Mittelwert hat. Man beachte jedoch, dass wir zu Y nur einen einzigen Wert beobachtet haben, der sich allerdings aus Mittelwerten zusammensetzt - Y wird daher bereits kleine Streuung haben.

Bemerkung zu sinnvoller Verallgemeinerung der Hypothesenformulierung: Sachlich ist es in aller Regel völlig uninteressant, etwa statistisch festzustellen, eine Studentenpopulation erbringe höhere Leistungen als eine andere. Vielmehr wird man auf eine Aussage der Form hinauswollen: „Gruppe B ist im Mittel um mehr als die Notendifferenz c im Mittel schlechter als Gruppe A", wobei c eine sachlich interessierende Differenz bedeutet. Diese Aussage mit gewisser Sicherheit behaupten zu können, das heißt, die Verneinung H_0: „Gruppe B ist im Mittel um höchstens c besser als A" auf einem guten Niveau α verwerfen zu können. H_0 bedeutet, wenn wir mit X die Variable „Note" bezeichnen: „$\mu(X_{\mid A}) - \mu(X_{\mid B}) \leq c$", also „$\mu(Y) \leq c$" ($c > 0$, und geringere Notenwerte sind die besseren). Dies bedeutet nun keinerlei neues technisches Problem, es macht nichts aus, ob eine Hypothese „$\mu(Y) \leq 0$" oder „$\mu(Y) \leq c$" mit einem beliebigen Wert c lautet. Es ist eine furchtbare Krankheit, diesen simplen Sachverhalt nicht zu sehen und dann lauter sachlich nichtssagende Hypothesen zu testen, wozu auch die unselige Bezeichnung von H_0 als „Nullhypothese" verführen mag. Weitere Verwirrung wird dadurch gestiftet, dass „signifikant" - auf Deutsch: „bedeutsam" *zwei verschiedene Bedeutungen* in unserem Kontext besitzt, einmal: „statistisch bedeutsam" im Sinne dessen, dass eine beobachtete Abweichung mit hoher Wahrscheinlichkeit nicht auf Zufall beruht, sodann „inhaltlich bedeutsame" Differenz. Es ist eine Unsitte, lediglich auf die erstere Bedeutung zu sehen und gar noch zu meinen, sie schließe letztere ein: Man kann völlig nichtssagende Behauptungen mit hoher statistischer Signifikanz versehen, aber man sollte sich selbstverständlich bemühen, inhaltlich wichtige Aussagen statistisch zu erhärten. Wir formulieren daher die benötigte Verallgemeinerung noch einmal ausdrücklich, die insbesondere für einseitige Fragestellungen von Bedeutung ist und die eben den Einbau inhaltlicher Bedeutsamkeit erlaubt:

	Hypothese (einseitig)
Hypothese	„$\mu(X_{\mid A}) \leq \mu(X_{\mid B}) + c$"
gleichwertige Umformulierung	„$\mu(\overline{X_{\mid A}} - \overline{X_{\mid B}}) \leq c$"

Grundlegend für das Rechnen sind nun folgende Feststellungen:

- $Y = \overline{X_{|A}} - \overline{X_{|B}}$ ist näherungsweise normalverteilt (wegen des Satzes, dass Summen unabhängiger normalverteilter Variablen normalverteilt sind), da $\overline{X_{|A}}$ und $\overline{X_{|B}}$ unabhängig und (zumindest näherungsweise) normalverteilt sind.
- $$\begin{aligned} \mu(Y) &= \mu\left(X_{|A}\right) - \mu\left(X_{|B}\right). \\ \sigma(Y) &= \sqrt{\frac{\sigma^2\left(X_{|A}\right)}{n_A} + \frac{\sigma^2\left(X_{|B}\right)}{n_B}}. \end{aligned}$$

  Dies folgt sofort aus der Linearität von μ und den Varianzformeln für $\overline{X}$ und für (linear) unabhängige Summen. Weiter folgt, dass man mit den Stichproben folgenden Schätzwert für $\sigma(Y)$ bei (normalerweise) unbekannten $\sigma\left(X_{|A}\right)$ und $\sigma\left(X_{|B}\right)$ hat:

  $$s(Y) = \sqrt{\frac{s^2\left(X_{|A}\right)}{n_A} + \frac{s^2\left(X_{|B}\right)}{n_B}},$$

  mit den oben eingeführten Schätzwerten für $\sigma^2\left(X_{|A}\right)$ und $\sigma^2\left(X_{|B}\right)$.

Bemerkung: In der Literatur findet man - leider - vielfach statt des letzteren $s(Y)$ einen anderen Streuungsschätzer, der nach bestandenem Test der Hypothese „$\sigma\left(X_{|A}\right) = \sigma\left(X_{|B}\right)$" durch Zusammenwerfen der Stichproben für eine gemeinsame Streuungsschätzung produziert wird. Anschließend wird ein $t-$ Test mit Einsetzen dieses Schätzers durchgeführt, analog dem oben beschriebenen für einfache Stichproben. Dies Verfahren ist jedoch nicht nur logisch völlig unsauber (man nimmt die Hypothese „$\sigma\left(X_{|A}\right) = \sigma\left(X_{|B}\right)$" als wahr an, nur weil man sie nicht hat verwerfen können), es liefert auch gerade in den Fällen kleinerer Stichprobenumfänge häufig völlig falsche Resultate (wie man mittels kleiner Computerexperimente leicht nachweist), wenn die Streuungen tatsächlich verschieden sind - und damit ist durchaus zu rechnen; es ist z.B. ganz typisch, dass eine Teilpopulation mit einem höheren $X-$ Mittelwert auch eine höhere Streuung aufweist. Daher werden wir dies ganze Verfahren mit keiner Silbe weiter beschreiben, das vielfach noch gar als einzige Methode („der $t-$ Test") erwähnt wird (!) und kompliziert, schlecht und völlig überflüssig ist. Wir werden stattdessen eine sehr bequeme und etwas ungenauere sowie eine immer noch recht bequeme und wirklich genaue Lösung des Problems angeben, die beide durchaus in guter Literatur bekannt sind.

2.3.1. *Erstes (ungenaueres) Verfahren: Anwenden der Normalverteilung.* Dies Verfahren zeigt brauchbare (bei hohen Stichprobenumfängen ausgezeichnete) Resultate, sobald die Stichprobenumfänge nicht zu klein sind (beide über 40 als grobe Faustregel). Zum Test der Hypothese, die wir stets in der Form „$\mu(Y) \leq [=]0$" formulieren können, auf Niveau α hat man folgende Vertrauensintervalle:

Hypothese	„$\mu(Y) \leq c$"	„$\mu(Y) = c$"
Vertrauensintervall	$c + z_{1-\alpha} s(Y)$	$c \pm z_{1-\alpha/2} s(Y)$

Wie zuvor ist auf Niveau α zu verwerfen, wenn der beobachtete Wert $y = \overline{x_A} - \overline{x_B}$ der Mittelwertdifferenzen außerhalb des betreffenden Vertrauensintervalls liegt. (Ebenso kann man wieder die entsprechende Gleichung nach z auflösen, um das Niveau zu bestimmen, auf dem man bei gegebener Beobachtung verwerfen könnte.)

Beispiel: H_0: „Gruppe A von Jugendlichen sieht im Mittel höchstens 1/2 Stunde täglich länger fern als Gruppe B" ($A \cap B = \varnothing$). Bezeichnen wir mit X

die Variable der mittleren täglichen Fernsehzeit für jede Person. Nun möge man beobachtet haben: $\overline{x_A} = 2.5$ Stunden, $s_A = 3/4$ Stunde, bei $n_A = 80$, $\overline{x_B} = 1.8$ Stunden, $s_B = 1/3$ Stunde, bei $n_B = 100$. Wir berechnen

$$s(Y) = s(\overline{X_{|A}} - \overline{X_{|B}}) = \sqrt{\frac{9/16}{80} + \frac{1/9}{100}} = 0.09$$

und erhalten die Gleichung

$$\frac{1}{2} + z\sqrt{\frac{9/16}{80} + \frac{1/9}{100}} = 0.7 (= \overline{x_A} - \overline{x_B}).$$

Das ergibt $z = 2.216\,4$, und $\alpha = \Phi_{0,1}(-2.21649) = 0.0133$, man könnte also auf 5%-Niveau verwerfen, hat jedoch das 1%-Niveau knapp verfehlt. Das bedeutet: Man ist recht sicher, dass Gruppe A im Mittel täglich mehr als 1/2 Stunde länger fernsieht als Gruppe B.

2.3.2. *Genaueres Verfahren: Anwenden der $t-$ Verteilungen.* Es ist tatsächlich kaum etwas zu ändern, lediglich ist eine etwas verwickelte Formel für die korrekte df = Freiheitsgrade zu verwenden, die im allgemeinen keine ganze Zahl mehr ergibt. Natürlich macht das kein Problem: Die zugehörigen $t-$ Verteilungen existieren, und man kann bei Benutzung von Tabellen die nächstkleinere ganze Zahl von Freiheitsgraden nehmen. Mit dem Computer hat man keinerlei Mehraufwand, da ein ordentliches Programm diese verallgemeinerten $t-$ Verteilungen besitzt.

Hier die Formel zur Berechnung der Freiheitsgrade; dabei bedeuten $n_{A,B}$ wie oben eingeführt die Stichprobenumfänge und sind mit $s_{A,B}$ die Streuungsschätzungen $s\left(X_{|A}\right), s\left(X_{|B}\right)$ abgekürzt - man beachte, dass man ohnehin $n_A, n_B > 1$ haben muss, weil es sonst keine Streuungsschätzungen $s_{A,B}$ gibt:

$$df = \frac{\left(s_A^2/n_A + s_B^2/n_B\right)^2}{\left(s_A^2/n_A\right)^2/(n_A - 1) + \left(s_B^2/n_B\right)^2/(n_B - 1)}.$$

Die Formel bewirkt, dass für $n_A = n_B$ und $s_A = s_B$ herauskommt: $df = n_A + n_B - 2$. Für sehr verschiedene Umfänge oder auch allein bei sehr verschiedenen $s_{A,B}$ kommt man dagegen kaum über den geringeren der beiden Stichprobenumfänge hinaus. Mit der so ausgerechneten Zahl df hat man dann folgende Vertrauensintervalle:

Hypothese	„$\mu(Y) \leq c$"	„$\mu(Y) = c$"
Vertrauensintervall	$c + t^{df}_{1-\alpha} s(Y)$	$c \pm t^{df}_{1-\alpha/2} s(Y)$

.

Beispiel: Wir behandeln dasselbe Fernsehdauer-Beispiel wie oben mit der verfeinerten Methode: Zusätzlich haben wir nur zu berechnen:

$$df = \frac{(9/(16 \cdot 80) + 1/(9 \cdot 100))^2}{(9/(16 \cdot 80))^2/79 + (1/(9 \cdot 100))^2/99} = 103.87.$$

Nunmehr ist die Wahrscheinlichkeit rechtsseitig von $t^{103.87} = 2.2164$ (die Gleichung bleibt dieselbe!) bzw. linksseitig von -2.2164 zur $t-$ Verteilung mit 103.87 Freiheitsgraden zu bestimmen, das ist etwa 0.144. Wie für solche Stichprobenumfänge versprochen, ist der Unterschied zum Resultat mit der einfachen Methode klein (das war 0.133). Aber dies (etwas ungünstigere) Ergebnis ist korrekter, und bei wesentlich kleineren Stichprobenumfängen können die Unterschiede drastisch werden.

KAPITEL 4

Chi-Quadrat (χ^2-)Tests

1. χ^2- Test der Unabhängigkeit von Merkmalen

Wenn man nur grob erst einmal wissen will, ob irgendwelche Merkmale überhaupt irgendetwas miteinander zu tun haben (nicht etwa denke man gleich an eines als Ursache des anderen!), so kann man nach Beobachtung der Merkmalsverteilung (genauer hier: der Häufigkeiten des Auftretens gewisse Kombinationen) in einer Stichprobe nach Wahrscheinlichkeit beurteilen, ob die Merkmale in der Gesamtbevölkerung noch unabhängig sein können. Hier lautet die zu testende Hypothese stets:

H_0 : „Die Merkmale ... sind unabhängig".

Das Schema des Testens ist (wie versprochen) dasselbe wie bei Hypothesen über unbekannte Mittelwerte. Allerdings folgen die Testvariablen einem anderen Verteilungstyp, der jedoch wiederum tabelliert ist. Man fragt also, ob man die Unabhängigkeitshypothese auf einem vorgegebenen Signifikanzniveau α verwerfen könne. Fällt der beobachtete Wert der Testvariablen in den Verwerfungsbereich, so kann man also mit der Sicherheit $1-\alpha$ behaupten, dass die Merkmale abhängig seien. (Allerdings ist das nur eine sehr grobe und schwache Aussage, später lernen wir verfeinertes zu den Stichworten „Korrelation" und allgemeiner: „Prädiktoren und Zielvariablen, lineare und andere Modelle".)

Das Schema des üblichen χ^2-Tests ist gerüstet für folgende

Grundsituation:

Man hat zwei Merkmale A, B, entweder einfache Eigenschaften, die nur zutreffen oder nicht zutreffen können - dann hat man nur $A, \overline{A}, B, \overline{B}$, oder allgemeiner jeweils mit mehreren (endlich vielen) Ausprägungen, dann hätte man etwa $A_1, ..., A_m, B_1, ..., B_n$. Einfaches Beispiel: Einkommen niedrig - mittel - hoch, Schulbildung gering - mittel - hoch. Es kann sich also durchaus um Vergröberungen von feineren quantitativen Merkmalen (Variablen) handeln, deren Ausprägungen durch Zahlenwert beschrieben werden (analog unseren Gruppierungen bei den Histogrammen), so dass man nach dieser Diskretisierung in jeweils nur endlich viele Ausprägungen den Test auch bei stetig verteilten Variablen mittelbar benutzen kann. Für den einfachen Fall mit je nur zwei Ausprägungen haben wir bereits Fisher's exakten Test kennengelernt: Dieser eignet sich jedoch nur für kleine absolute Häufigkeiten, während es sich beim χ^2- Test um eine Näherung der eigentlich diskreten hypergeometrischen Verteilung handelt, welche gerade für nicht zu kleine Einträge gut funktioniert!

Datenstruktur der in einer Stichprobe beobachteten Daten

Wenn wir eine Stichprobe vom Umfang N aus der Gesamtpopulation bilden und festhalten, welche Merkmalskombinationen $A_i \cap B_j$ mit welcher *absoluten Häufigkeit* (diese nimmt man hier!) vorkommen, so ergibt sich bei sinnvoller Anordnung automatisch eine Tafel der folgenden Art:

($m \times n$)-Felder-Tafel (mit zusätzlichen Beschriftungen und Randsummen

Auspr. von $B \rightarrow$ Auspr.von A $\downarrow$	B_1	$\cdots$	B_n	Zeilensummen $\downarrow$
A_1	$\underbrace{f_{11}}_{\text{Anzahl in } A_1 \cap B_1}$	$\cdots$	f_{1n}	$a_1 = \sum_{j=1}^{n} f_{1j}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_m	f_{m1}	$\cdots$	$\underbrace{f_{mn}}_{\text{Anzahl in } A_m \cap B_n}$	$a_m = \sum_{j=1}^{n} f_{mj}$
Spaltensummen $\rightarrow$	$b_1 = \sum_{i=1}^{m} f_{i1}$	$\cdots$	$b_n = \sum_{i=1}^{m} f_{in}$	$N = \sum_{i=1}^{m} a_m = \sum_{j=1}^{n} b_n$

Dabei sind die Einträge f_{ij} die absoluten Häufigkeiten der Objekte in der Stichprobe (vom Umfang N), welche in $A_i \cap B_j$ liegen, also die Ausprägung A_i des Zeilenmerkmals und B_j des Spaltenmerkmals haben. a_i ist die Anzahl der in A_i gefundenen Objekte, daher ergibt sich a_i als Summe der f_{ij} über j, und daher ist N die Summe der a_i. Analog ist b_j die Anzahl der in B_j gefundenen Objekte, also die Summe der f_{ij} über i, und die Summe der b_j ist wieder N.

Beispiel: $n = m = 2$ (also eine (2×2)-Felder-Tafel, $A_1 = A$, $A_2 = \overline{A}$, $B_1 = B$, $B_2 = \overline{B}$):

	B	$\overline{B}$	
A	60	30	90
$\overline{A}$	40	90	130
	100	120	220

Hier hätte man also: $f_{11} = 60$, $f_{12} = 30$, $f_{21} = 40$, $f_{22} = 90$, $a_1 = 90$, $a_2 = 130$, $b_1 = 100$, $b_2 = 120$, $N = 220$.

Das Problem: Sind Zeilen- und Spaltenmerkmal unabhängig voneinander? Hypothese H_0: „Sie sind es".

Diese Hypothese wird einem statistischen Test unterworfen, dem χ^2 -Test. Er ist so benannt wegen der (ungefähren, idealisierten) χ^2 -Verteilung der dabei benutzten Prüfvariablen oder Testvariablen. Diese χ^2 -Verteilungen bilden wieder eine ganze Familie, ähnlich wie bei den t-Verteilungen ist es eine Familie mit einem Parameter, der wiederum ganzzahlig ist, ab $n = 1$, und ebenfalls „Zahl der Freiheitsgrade" heißt.

Ausführlicher: Hat man keine veränderte Wahrscheinlichkeit für irgendeine Ausprägung von B [bzw. von A] , wenn irgendeine bestimmte Ausprägung von A [bzw. von B] vorliegt? Diese Frage stellt man für die Gesamtpopulation. Natürlich wird es auch dann, wenn die Unabhängigkeitshypothese wahr ist, in der Stichprobe eine leichte Abweichung von der Unabhängigkeit geben. Wie nunmehr schon gewohnt schauen wir nach, ob diese Abweichung signifikant ist, d.h. zu groß, um noch als Zufallsabweichung durchgehen zu können. Wie stünde das im gegebenen $(2 \times 2)-$ Beispiel? Dazu müssen wir zunächst einmal ein vernünftiges Maß für die Abweichung von der Unabhängigkeit im Rahmen der gefundenen Stichprobe bilden. Was taten wir beim Test einer Hypothese über einen Mittelwert, insbesondere eine relative Häufigkeit? Wir verglichen den hypothetischen Wert mit dem in der Stichprobe beobachteten. In unserer neuen Situation bekommen wir ein Analogon,

eine Maßzahl für die Abweichung, auf folgende Weise: Wir haben *alle* Einträge der Matrix (Tafel) mit den laut Unabhängigkeitshypothese erwarteten Werten zu vergleichen, testen also so etwas wie den Unterschied mehrerer Mittelwerte von hypothetischen. Da liegt es nahe, die einzelnen (im Beispiel sind es 4) Differenzen zu quadrieren und diese Quadrate dann zu addieren. Nun werden mit dem Stichprobenumfang auch die Differenzen erwartungsgemäß größer, und daher müssen wir, um vergleichbare Maßzahlen zu bekommen, eine Normierung vornehmen. Wahrscheinlichkeitstheoretisch günstig (so dass man die resultierende Verteilung gut kennt) ist es, die einzelnen quadrierten Differenzen noch durch die erwartete Häufigkeit des zugehörigen Feldes zu teilen. Dann haben wir (mit der Summe der normierten Quadrate) eine einzige Zahl, die das Maß der Abweichung beschreibt. Nun wollen wir noch einsehen, auf welche Weise die Unabhängigkeitshypothese tatsächlich erwartete Werte für die Tafelfelder nach sich zieht. Wir illustrieren das zunächst am konkreten Beispiel, formulieren es nachher dann für die abstrakte, allgemeine $(m \times n)-$ Tafel. In einem ersten Schritt setzen wir die Randsummen geteilt durch Stichprobenumfang, also die *beobachteten relativen Häufigkeiten der (im Beispiel einfachen - es gibt nur „ja"/„nein") Merkmale* A, B als *Schätzwerte* für die (unbekannten) relativen Häufigkeiten dieser Merkmale in der Gesamtpopulation an. (Für die Komplemente $\overline{A}$ und $\overline{B}$ haben wir dann automatisch Schätzwerte, die wir aber auch entsprechend über Randsummenwert$/N$ gewinnen könnten). Zusammen für die Beispieltafel:

$p_A = a_1/N = 90/220 = 9/22$, damit $p_{\overline{A}} = a_2/N = 13/22$.

$p_B = b_1/N = 100/220 = 5/11$, damit $p_{\overline{B}} = 6/11$.

Was sagt nun die Unabhängigkeitshypothese? Sie sagt zunächst

$$P(A \cap B) = P(A) \cdot P(B).$$

Aber daraus folgt sofort:

$$P(A \cap \overline{B}) = P(A) \cdot P(\overline{B}),\ P(\overline{A} \cap B) = P(\overline{A}) \cdot P(B),\ P(\overline{A} \cap \overline{B}) = P(\overline{A}) \cdot P(\overline{B}).$$

Somit sagt die Unabhängigkeitshypothese all dies. Das bedeutet aber, dass aus unseren Schätzwerten für $P(A), P(B)$ unter Voraussetzung dieser Hypothese (man erinnere sich an die Logik des Hypothesentestens: die Hypothese wird vorausgesetzt, und es wird daraus ein Vertrauensintervall für die jeweilige Testvariable abgeleitet) Schätzwerte für die Wahrscheinlichkeiten für jedes Feld resultieren, und daraus erhalten wir durch Multiplikation mit N erwartete Einträge (absolute Häufigkeiten) für alle Felder. Im Beispiel ergibt sich folgende Tafel der erwarteten Häufigkeiten:

	B	$\overline{B}$	
A	$220 \cdot (9/22) \cdot (5/11) = 450/11$	$220 \cdot (9/22) \cdot (6/11) = 540/11$	90
$\overline{A}$	$220 \cdot (13/22) \cdot (5/11) = 650/11$	$220 \cdot 13/22 \cdot 6/11 = 780/11$	130
	100	120	220

(Die Randsummen bleiben dieselben, die Erwartungswerte sind naturgemäß im allgemeinen krumme Zahlen.) Nun bilden wir unsere Maßzahl für die Abweichung des Beobachteten von der Unabhängigkeit und nennen sie χ_1^2 (als beobachteten Wert einer χ^2 -verteilten Variablen mit einem Freiheitsgrad):

$$\chi_1^2 = \frac{\left(60-\frac{450}{11}\right)^2}{\left(\frac{450}{11}\right)} + \frac{\left(30-\frac{540}{11}\right)^2}{\left(\frac{540}{11}\right)} + \frac{\left(40-\frac{650}{11}\right)^2}{\left(\frac{650}{11}\right)} + \frac{\left(90-\frac{780}{11}\right)^2}{\left(\frac{780}{11}\right)} = 27.641$$

Erläuterung der „Freiheitsgrade"

Wir erklären, wieso hier ein Freiheitsgrad vorliegt: Das meint, dass nur einer dieser vier Summanden unabhängig wählbar ist, dass man also alle anderen daraus berechnen kann. Dies ist so einzusehen: Wir haben die Randsummen fixiert (zur Schätzung der Wahrscheinlichkeiten der Zeilen- und Spaltenmerkmale). Damit liegen laut Unabhängigkeitshypothese alle erwarteten Einträge fest. Wenn man nun den ersten Summanden (für das Feld links oben) von χ_1^2 kennt, dann lässt sich daraus der Eintrag 60 der Beobachtungstafel rekonstruieren. Daraus lassen sich aber bei den fixierten Randsummen die restlichen Einträge der Beobachtungstafel ausrechnen und damit die letzten drei Summanden des Wertes χ_1^2. (Auch bei der t-Verteilung gab die Zahl der Freiheitsgrade die Zahl der unabhängigen Summanden einer Summe von (normierten) Quadraten an!) Bei einer höheren Zahl an unabhängigen Summanden sollte man auch im Mittel und mit größeren Wahrscheinlichkeiten höhere Werte erhalten. Daher hat man die verschiedenen χ^2 -Verteilungen für die einzelnen Freiheitsgrade. Wie sähe das aus bei einer allgemeinen $(m \times n)$-Tafel? Man sieht sofort, dass bei festgelegten Randsummen $(m-1)\cdot(n-1)$ Feldereinträge in der beobachteten Tafel frei vorkommen können. Entsprechend sind aus den zugehörigen χ^2 -Summanden die restlichen rekonstruierbar. Daher haben wir das allgemeine Resultat:

Für eine $(\mathbf{m \times n})$-Tafel hat die Maßgröße $\boldsymbol{\chi}^2$ für die Abweichung von der Unabhängigkeitshypothese $\mathbf{(m-1)\cdot(n-1)}$ Freiheitsgrade.

Wie ist nun die Unabhängigkeitsfrage im Beispiel zu beantworten? Wir schauen einmal in der Tafel zu den χ^2 -Verteilungen nach, mit welcher Wahrscheinlichkeit ein Wert ≥ 27.641 bei einer χ^2 -Verteilung mit einem Freiheitsgrad auftritt: Wir sehen gerade noch, dass bereits Werte über 10.83 nur noch mit Wahrscheinlichkeit 1/1000 vorkommen. In unserem Beispiel ist also die Abweichung hochsignifikant - d.h. die Unabhängigkeitshypothese ist mit überwältigender Sicherheit zu verwerfen, wie man anhand der Zahlen wohl auch vermutet hätte.

Wir haben der Einfachheit halber die allgemeine Sache am $(2 \times 2)-$ Beispiel erklärt. Man sollte mittels dieses Beispiels leichter nachvollziehen, was bei größeren als $(2 \times 2)-$ Tafeln zu tun ist, für die Zahl der Freiheitsgrade s.u. Abschnitt 1.2.) Nun bildet (2×2) tatsächlich in doppelter Hinsicht eine Ausnahme, so dass man die obenstehenden Ausführungen gerade für diesen Fall nicht als rezeptartige Beschreibung des Vorgehens missverstehen sollte. Erstens ist die Sache rechentechnisch für $(2 \times 2)-$ Tafeln bedeutend einfacher zu handhaben, d.h. man kann den Wert χ_1^2 viel leichter ausrechnen als oben geschehen (für größere Tafeln dagegen muss man rechnen wie im Beispiel gezeigt!). Zweitens ergibt sich eine Komplikation (die jedoch rechentechnisch sehr einfach zu handhaben ist): Man berechnet viel genauere Wahrscheinlichkeiten, wenn man eine Stetigkeitskorrektur durchführt (sog. Yates-Korrektur). Wie man mit dem Sonderfall (2×2) praktisch umgeht, wird in 1.2 gezeigt.

1.1. Der Fall aller Tafeln größer als (2×2). Zur allgemeinen beobachteten $(m \times n)-$ Tafel aus 1. weiter oben bestimmt man mittels der beobachteten Randhäufigkeiten

1. Geschätzte Randwahrscheinlichkeiten

$$p_i = \frac{a_i}{N},\ \text{(Schätzwert für } P(A_i)),\ 1 \le i \le m,$$

$$q_j = \frac{b_j}{N},\ \text{(Schätzwert für } P(B_j)),\ 1 \le j \le n.$$

Daraus ergeben sich mittels der Unabhängigkeitshypothese geschätzte Wahrscheinlichkeiten dafür, dass ein Populationsmitglied in $A_i \cap B_j$ liegt:

$$r_{ij} = p_i \cdot q_j\ \text{(Schätzwert für } P(A_i \cap B_j)),\ 1 \le i \le m,\ 1 \le j \le n.$$

Daraus erhält man:

2. Laut Unabhängigkeitshypothese erwartete (absolute) Häufigkeiten für die Tafelfelder, für das Feld der i. Zeile und der j. Spalte:

$$e_{ij} = r_{ij} \cdot N = \frac{a_i \cdot b_j}{N},\ 1 \le i \le m,\ 1 \le j \le n.$$

3. Berechnung von $\chi^2_{(m-1)\cdot(n-1)}$:

$$\chi^2_{(m-1)\cdot(n-1)} = \sum_{\substack{1\le i\le m\\ 1\le j\le n}} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}.\ \text{Dazu müssen alle } e_{ij} \neq 0 \text{ sein!}$$

4. Die Unabhängigkeitshypothese wird verworfen auf dem Signifikanzniveau α genau dann, wenn ein Wert einer $\chi^2_{(m-1)\cdot(n-1)}$ -verteilten Größe mit $(\mathbf{m-1})\cdot(\mathbf{n-1})$ Freiheitsgraden einen so hohen oder höheren Wert nur mit einer Wahrscheinlichkeit $< \alpha$ annimmt. Das entnimmt man einer Tabellierung der χ^2 -Verteilungen.

Beispiel:

Beobachtete Tafel sei

elterlicher Erziehungsstil → Straffälligkeit ↓	liberal	autoritär	laissez faire	
straffällig	10	30	50	90
nicht straffällig	2000	2500	4000	8500
	2010	2530	4050	8590

Wir zeigen das Resultat vom 2. Schritt (den man wie illustriert sofort vornehmen kann) in einer Tafel der erwarteten absoluten Häufigkeiten (gemäß Unabhängigkeitshypothese):

Erz.-Stil → Straffälligkeit ↓	liberal	autoritär	laissez faire	
straffällig	$90 \cdot 2010/8590$	22770/859	36450/859	90
nicht straffällig	854250/431	2150500/859	3442500/859	8500
	2010	2530	4050	8590

Dies ergibt (3. Schritt) (natürlich würde man mit Dezimalzahlen im Taschenrechner arbeiten und die einzelnen Werte nicht erst aufschreiben, sondern im Rechner lassen):

$$\frac{(10-\frac{18090}{859})^2}{\frac{18090}{859}}+\frac{(30-\frac{22770}{859})^2}{\frac{22770}{859}}+\frac{(50-\frac{36450}{859})^2}{\frac{36450}{859}}+\frac{(2000-\frac{854250}{431})^2}{\frac{854250}{431}}+\frac{(2500-\frac{2150500}{859})^2}{\frac{2150500}{859}}+\frac{(4000-\frac{3442500}{859})^2}{\frac{3442500}{859}}=$$

$$\chi_2^2 = 7.7997,$$

und diese Zahl ist so groß, dass die Wahrscheinlichkeit eines so hohen oder höheren χ^2 -Wertes bei 2 Freiheitsgraden weniger als 2.5% beträgt. Also ist die Unabhängigkeitshypothese auf dem 2.5%-Niveau zu verwerfen. (Schon aus der Beobachtungstabelle hätte man ersehen sollen, dass insbesondere die Kinder von „laissez-faire"-Eltern bei den Straffälligen überrepräsentiert waren. Die Frage ist halt immer nur, ob das Ausmaß ausreicht, um auf die Gesamtpopulation zu schließen. Im Beispiel war das so.)

1.2. Das Vorgehen bei (2×2)-Tafeln. Das Verfahren ist gegenüber dem von 1.1 für größere Tafeln stark zu vereinfachen:

Gegeben sei die beobachtete Tafel (mit vereinfachter Notation unter Vermeidung der doppelten Indizes, die hier nicht lohnen):

	B	$\overline{B}$	
A	a	b	$m_1 = a+b$
$\overline{A}$	c	d	$m_2 = c+d$
	$n_1 = a+c$	$n_2 = b+d$	$N = m_1+m_2 = n_1+n_2$

Dann haben wir folgendes

Resultat:

Die χ_1^2 -Summe aus den vier Summanden, wie oben berechnet, lässt sich durch folgende Formel exakt und viel leichter berechnen:

$$\chi_1^2 = \frac{(ad-bc)^2 \cdot N}{m_1 \cdot m_2 \cdot n_1 \cdot n_2}.$$

Aber diese Formel sollte man so nicht benutzen und in die Tabelle gehen. Der Grund ist folgender: Die χ^2 -Verteilung mit einem Freiheitsgrad ist eine stetige, mathematisch idealisierte Verteilung (darin ähnlich den Normalverteilungen). Aber zumal bei kleineren Randsummen treten nur sehr wenige überhaupt beobachtbare Werte χ_1^2 auf, und tatsächlich kommt dabei - und zwar auch noch bei großen N und Randsummen! - ein Bias heraus: Die Wahrscheinlichkeiten für mindestens so große χ_1^2 -Werte werden systematisch *unterschätzt*, wie man unter Nutzung von Fisher's exaktem Test zeigen kann (s.u. 1.3). Das geht genau nach der falschen Seite: Man verwirft mit einer vorgeblichen Sicherheit, die gar nicht da ist. Daher bringt man (ganz ähnlich wie bei der Näherung von Binomialverteilungen durch Normalverteilungen) eine Stetigkeitskorrektur an, die darauf hinausläuft, in geeigneter Weise den χ_1^2 -Wert zu reduzieren. Dabei zeigt sich (vgl. 1.3), dass man zu hervorragenden Näherungen der tatsächlichen kombinatorischen Wahrscheinlichkeiten gelangt. Die Frage liegt nahe, ob eine analoge Korrektur nicht auch für größere Tafeln vorzunehmen wäre. Es zeigt sich jedoch im Computerexperiment, dass man mit den unkorrigierten Werten selbst bei recht kleinen Einträgen (Regeln wie die, dass keiner unter 5 liegen sollte, erweisen sich *dafür* als übervorsichtig) sehr gute Ergebnisse erhält, und dass verschiedenste Korrekturversuche, auch noch so vorsichtige, nur in Einzelfällen und nicht systematisch Verbesserungen bringen. Man

sollte sie daher bei größeren Tafeln unterlassen. Dagegen sollte die folgende Stetigkeitskorrektur für $(2 \times 2)-$ Tafeln *stets* vorgenommen werden - sie bringt beinahe immer deutlich bessere und niemals schlechtere Resultate, und sie verlangt keinen höheren Aufwand.

Stetigkeitskorrektur (nur) für χ_1^2 bei $(2 \times 2)-$ Tafeln

1. Schritt: Man schaue nach, ob $|ad - bc| \geq N/4$ (Absoluter Betrag). Nur dann nehme man folgende Korrektur (2. Schritt) vor.

(Im andern Fall ist nichts verloren, dann wird man eh kaum geneigt sein, noch zu rechnen, weil die Abweichung viel zu klein sein wird, um auf verbünftigem Niveau die Unabhängigkeitshypothese zu verwerfen.) Es sei bemerkt, dass man mit der folgenden Korrektur ohne Beachtung dieser Bedingung genau das Gegenteil dessen produzieren kann, was man beabsichtigt. Ist die Bedingung nicht erfüllt, so wird nämlich der „korrigierte" χ^2 -Wert noch höher als der unkorrigierte, das Resultat noch falscher. Es ist schon eine gewisse Ironie, dass die Yatessche Korrektur um 1950 von Leuten öffentlich angegriffen wurde, die zu dämlich waren, das zu beachten.

2. Schritt: Man verändere die Einträge a,b,c,d der Tafel wie folgt:

Die folgenden etwas pedantischen Ausführungen sind im Grunde überflüssig, man kann die Sache ganz knapp so formulieren: Man schaue die Diagonalenprodukte ad, bc an. Im größeren der beiden setze man die einzelnen Einträge (also a, d oder aber c, d) um $1/2$ herab, im kleineren um $1/2$ herauf. Nun die ausführlicher Form:

1. Fall: $ad > bc$. Dann setze man

$$\begin{aligned} a' &= a - 0.5 \\ d' &= d - 0.5 \\ b' &= b + 0.5 \\ c' &= c + 0.5 \end{aligned}$$

(mit dem Effekt, dass der Wert $(ad - bc)^2$ vermindert wird, falls die im ersten Schritt geprüfte Bedingung zutrifft).

2. Fall: $bc > ad$. Dann setze man

$$\begin{aligned} a' &= a + 0.5 \\ d' &= d + 0.5 \\ b' &= b - 0.5 \\ c' &= c - 0.5 \end{aligned}$$

(mit demselben Effekt, dass der Wert $(ad - bc)^2$ vermindert wird, falls die im ersten Schritt geprüfte Bedingung zutrifft).

3. Schritt: Man rechne den korrigierten Wert $\chi^2_{1,korr}$ mit der Tafel der korrigierten Einträge genau nach der vereinfachten Formel aus, also

$$\chi^2_{1,korr} = \frac{(a'd' - b'c')^2 \cdot N}{m_1 \cdot m_2 \cdot n_1 \cdot n_2}.$$

Exkurs zu einer gleichwertigen Formel und zur Begründung des ersten Schrittes:

Die Formel

$$\chi^2_{1,korr} = \frac{\left(|ad-bc| - \frac{N}{2}\right)^2 \cdot N}{m_1 \cdot m_2 \cdot n_1 \cdot n_2}$$

rechnerisch dasselbe Resultat, aber das ist nicht bequemer, und die erste Formel lässt sich dem Sinne nach besser verstehen. Allerdings lässt die zweite Formel besser einsehen, dass man, um den gewünschten Effekt

$$\chi^2_{1,korr} \leq \chi^2_1$$

zu erhalten, gerade die Bedingung $|ad-bc| \geq N/4$ benötigt:

Wir nehmen einmal an, dass $ad > bc$ sei (sonst sind diese Produkte einfach zu vertauschen, und das Argument verläuft ebenso, wie gleich ausgeführt wird). Die Bedingung $\chi^2_{1,korr} \leq \chi^2_1$ bedeutet so viel wie

$$(a'd' - b'c')^2 \leq (ad-bc)^2.$$

Mit $ad > bc$ bedeutet das nach Definition:

$$\begin{aligned}\left(\left(a-\frac{1}{2}\right)\left(d-\frac{1}{2}\right) - \left(b+\frac{1}{2}\right)\left(c+\frac{1}{2}\right)\right)^2 &= \left(ad-bc-\frac{1}{2}(a+b+c+d)\right)^2 \\ &= \left(ad-bc-\frac{N}{2}\right)^2 \leq (ad-bc)^2.\end{aligned}$$

Damit haben wir bereits gezeigt, dass die beiden Formeln wirklich dasselbe Resultat liefern! Denn mit $ad > bc$ ist $ad - bc = |ad - bc|$. (Für $bc > ad$ hätten wir gerade bei b, c den Wert $1/2$ abgezogen, und es stünde $bc - ad - \frac{N}{2}$, was dann dasselbe wie $|bc - ad| - \frac{N}{2}$ wäre, und natürlich gilt stets $|ad - bc| = |bc - ad|$.)

Nun arbeiten wir mit der allgemeinen Form der zweiten Formel, welche die Betragstriche benutzt.

$$\left(|ad-bc| - \frac{N}{2}\right)^2 = (ad-bc)^2 - N|ad-bc| + \frac{N^2}{4}.$$

Dies ist $\leq (ad-bc)^2$ genau dann, wenn gilt

$$-N|ad-bc| + \frac{N^2}{4} \leq 0, \text{ also}$$

$$N(ad-bc) \geq \frac{N^2}{4} \text{ oder gleichwertig}$$

$$ad - bc \geq \frac{N}{4}, \text{ was mit } ad \geq bc \text{ dasselbe ist wie}$$

$$|ad-bc| \geq \frac{N}{4}.$$

4. Schritt: Wie im andern Fall, nur geht man mit diesem korrigierten Wert in die Tabelle und verwirft die Unabhängigkeitshypothese,

wenn die dem Niveau zugehörige Signifikanzgrenze mit ihm überschritten wird.

Beispiel:

	suchtgefährdet	nicht suchtgefährdet	
Männer	40	200	240
Frauen	20	180	200
	60	380	440

$$\chi^2_{1,korr} = \frac{(39.5 \cdot 179.5 - 20.5 \cdot 200.5)^2 \cdot 440}{240 \cdot 200 \cdot 60 \cdot 380} = 3.5703,$$

man könnte also nicht auf 5%-Niveau verwerfen, würde aber keineswegs an die Unabhängigkeit glauben, da man das Verwerfen auf 5%-Niveau nur sehr knapp verfehlt, 3.84 ist die kritische Grenze. Übrigens kann man mittels der Standard-Normalverteilung genau sagen, auf welchem Niveau man noch verwerfen könnte; denn eine χ^2_1-verteilte Variable ist das Quadrat einer standard-normalverteilten Variablen: Die Zahl 3.84 ist das Quadrat von der bestens vertrauten Zahl 1.96, die zum zweiseitigen 95%-Vertrauensintervall der Standard-Normalverteilung gehört. Wir bilden also $\sqrt{3.5703} = 1.8895$ und sehen, dass dazu die Wahrscheinlichkeit $0.9699 - (1 - 0.9699) = 0.9398$ gehört (beim Übergang zur Normalverteilung: zweiseitig! - Der χ^2 -Test ist dagegen *formal* stets einseitig, *inhaltlich* dagegen zweiseitig, s.u. 1.4). Wir hätten also auf beinahe 6%-Niveau verwerfen können. Vergleichen wir, was wir bekommen hätten mit unkorrigiertem χ^2_1-Wert: $\chi^2_1 = \frac{(40 \cdot 180 - 20 \cdot 200)^2 \cdot 440}{240 \cdot 200 \cdot 60 \cdot 380} = 4.117$. Hätten wir damit gearbeitet, so wären wir zu dem ziemlich falschen Resultat gelangt ,dass nicht nur auf 5%-Niveau, sondern beinahe sogar auf 4%-Niveau zu verwerfen wäre. Genauer hätte man damit auf dem Niveau 0.0424 (oder 4.24%) verwerfen können. (Das exakte Resultat liegt knapp über 0.05, vgl. die ausführlichere Besprechung diese Punktes im folgenden Abschnitt 1.3.) Wir sehen im Beispiel, wie unkorrigierte χ^2 -Werte nach der unsicheren Seite tendieren.

1.3. Zur Qualität der Näherung der exakten kombinatorischen Wahrscheinlichkeiten durch die χ^2-Verteilungen.

In den komplizierteren Fällen der größeren Tafeln als nur (2×2) funktioniert solche Anwendung ebenfalls, ist aber etwas schwieriger und wird hier ausgelassen bzw. zur Anregung als Problem gegeben. Wir denken uns die Situation zu einer beobachteten Tafel und nehmen erneut das Zahlenbeispiel aus Kapitel 2, 2.3 (S. 31) vor:

	B	$\overline{B}$	
A	6	5	11
$\overline{A}$	3	10	13
	9	15	24

.

Mittels Fisher's exakten Tests ergab sich als Signifikanzniveau $\alpha = 0.21$, auf dem in diesem Falle die Unabhängigkeitshypothese zu verwerfen ist, so dass man lieber nicht verwürfe. Wir fragen, wie genau die (wesentlich bequemere!) χ^2-Näherung in der Lage ist, diese Wahrscheinlichkeit α zu reproduzieren. Ohne Stetigkeitskorrektur erhalten wir

$$\chi^2_{1,unkorr} = 2.52, \text{ das heißt } \alpha = 0.112.$$

(in Ihrer Tafel sehen Sie immerhin, dass α zwischen 0.1 und 0.25 liegt, näher an 0.1 - mit der oben beschriebenen Reduktion auf die Standard-Normalverteilung bei einem Freiheitsgrad ersehen Sie genauer: $\alpha = 2 - 2\Phi_{0,1}(\sqrt{2.52}) = 0.112.$). Das ergibt also ein deutlich viel zu günstiges Niveau, obgleich es immer noch nicht zu recht sicherem Verwerfen reicht. *Mit Stetigkeitskorrektur* erhalten wir immerhin:

$$\chi^2_{korr} = 1.354, \text{ das heißt } \alpha = 0.245.$$

Das zeigt, wie viel besser die Sache mit Stetigkeitskorrektur wird, es zeigt aber auch, dass die Rechnung (wenn auch zur sicheren Seite!) immer noch etwas ungenau ist. Eine recht gute Regel ist es, χ^2 nicht anzuwenden, wenn mindestens ein Eintrag unter 5 liegt. (Bei größer dimensionierten Tafeln, also mindestens 3 Zeilen oder Spalten, wird die Näherung übrigens besser, und es gibt dann auch keine lohnende Stetigkeitskorrektur!) Spielen wir auch dies an einem Beispiel mit immer noch kleinen Einträgen durch:

	B	$\overline{B}$	
A	5	10	15
$\overline{A}$	10	5	15
	15	15	30

.

Hier ergibt Fisher's Test (exakte hypergeometrische Rechnung):

$$\alpha = 1 - 2\sum_{i=10}^{15} \frac{\binom{15}{i}\binom{15}{15-i}}{\binom{30}{15}} = 0.143.$$

χ^2- Test ohne Stetigkeitskorrektur ist wieder herzlich schlecht:

$$\chi^2_{unkorr} = 3.33, \text{ entsprechend } \alpha = 0.068.$$

Dagegen bemerken wir, dass Stetigkeitskorrektur folgendes sehr ordentliche Resultat liefert:

$$\chi^2_{korr} = 2.133, \text{ entsprechend } \alpha = 0.144.$$

Allgemein können wir sagen, dass auf die Stetigkeitskorrektur im $(2 \times 2)-$ Falle Verlass ist, wenn kein Eintrag unter 5 liegt. Das Unterlassen der Korrektur ist dagegen niemals ratsam, zumal da sie keine zusätzliche Mühe bereitet.

1.4. Einseitigkeit und Zweiseitigkeit bei χ^2-Unabhängigkeitstests. Man sollte sich klarmachen, dass die Unabhängigkeitshypothese von vornherein zweiseitig ist - wenn in der Stichprobe A in B unterrepräsentiert ist, so spricht das ebenso gegen diese Hypothese wie eine Überrepräsentierung von A in B. Tatsächlich erhält man mit der Tafel

5	10
10	5

denselben χ^2 -Wert wie mit der Tafel

10	5
5	10

. Das Quadrieren bewirkt, dass Abweichungen in beiden Richtungen zählen. So haben wir auch bei der exakten Rechnung jeweils beide Reihen der abweichenden Tafeln in beiden Richtungen zusammengenommen und alle zugehörigen Wahrscheinlichkeiten addiert. Aber manchmal hat man von vornherein eine einseitige Fragestellung, will etwa nur eine Hypothese der Form, dass B in A nicht häufiger vorkommt als in $\overline{A}$, verwerfen. Dazu erläutern wir noch, wie man für $(2 \times 2)-$ Tafeln mittels der χ^2 -Methode auch zugehörige einseitige Tests durchführen kann. Das Verfahren ist sehr einfach: Weicht die Tafel nicht in der relevanten Richtung vom Erwarteten ab

(ist also B in A in der Stichprobe nicht überrepräsentiert), so spricht das jedenfalls nicht gegen die Hypothese. Im gegenteiligen Falle jedoch berechnet man die Wahrscheinlichkeit einer mindestens so großen Abweichung wie beobachtet. Das heißt, man nimmt nur eine der Tafelreihen. Das ergibt eine Wahrscheinlichkeit, die etwa halb so groß ist wie die zweiseitig berechnete. (Bei gleichen Zeilen- oder Spaltensummen ist es genau die Hälfte.) Fazit: Man berechnet den korrigierten χ^2 -Wert wie bisher, verwirft aber bereits auf Niveau $\alpha/2$, wenn man zweiseitig auf Niveau α verwerfen könnte.

Beispiel: Hypothese: Das Lesehobby ist unter den Mädchen nicht häufiger als unter den Jungen. Beobachtet wurde:

	Lesen als Hobby	kein Lesehobby
Mädchen	50	40
Jungen	20	50

.

Es ist $\chi^2_{1,korr} = 10.58$, die Wurzel davon liegt zwischen 3.2 und 3.3, also ist die Wahrscheinlichkeit für kleinere χ^2 -Werte über 0.9993, unter 0.9995. Man kann also auf dem Niveau 7/10000 verwerfen, auf 5/10000 nicht mehr. (Wir sind einfach zur zugehörigen normalverteilten Variablen gegangen und haben das einseitige Intervall genommen.) Mit der χ^2 -Tabelle funktioniert die Sache so, dass man das abgelesene Niveau noch einmal halbiert. Man liest ab, dass man zweiseitig beinahe auf 1/1000 verwerfen könnte (dafür braucht man den χ^2 -Wert 10.83), also einseitig beinahe auf 5/10000. Das stimmt. Die exakte Rechnung ergibt übrigens in unserem Falle $\sum_{i=50}^{90} \frac{\binom{70}{i}\cdot\binom{90}{90-i}}{\binom{160}{90}} = 0.00051871$, also sehr geringfügig mehr als 5/10000.

2. χ^2- Anpassungstests

2.1. Fragestellung und Lösungsverfahren. Die Fragestellung lautet hier spezifisch, ob ein zufällig gezogener Datensatz von einer bestimmten Verteilung herrührt. Man hat zum Beispiel eine empirische Verteilung (zum Beispiel eine von Abschlussnoten) aus einer Stichprobe und prüft, ob sie einer bestimmten Verteilung (zum Beispiel Normalverteilung mit passenden Werten für μ, σ) entspricht. Aber was heißt „entsprechen"? Immer wird es in einer Stichprobe gewisse Abweichungen vom Erwarteten geben, auch wenn die Hypothese korrekt ist. Es kommt wie gewohnt darauf an, welche Abweichungen zu groß sind, um noch als Zufallsfluktuationen erklärt werden zu können.

Die neue Fragestellung ist *abstrakt gesehen* völlig analog zur Unabhängigkeitsfrage, die wir zuvor mittels χ^2- Tests behandelten: Auch dort war die Frage, ob die beobachteten Häufigkeiten einer bestimmten Verteilung entsprachen, nämlich den Produktwahrscheinlichkeiten aus den Randwahrscheinlichkeiten. Daher können wir auch die neue Fragestellung ganz analog mittels χ^2- Tests angehen: Zunächst haben wir (wie zuvor) dafür zu sorgen, dass wir den Bereich aller Variablenwerte gegebenenfalls in nur endlich viele Intervall-Ereignisse einteilen - eine Gruppierung ist übrigens auch dann sinnvoll und nützlich, wenn es gar zu viele einzelne Werte gibt, da sonst viele Werte in der Stichprobe gar nicht vorkommen werden. Anschließend werden wir für jedes dieser Intervalle (oder im Falle weniger Werte auch jeden einzelnen Wert) - also für jedes „Haus" - den Wert von χ^2 wie gewohnt ausrechnen als Summe der Quadrate der Differenzen zwischen beobachteten und (laut hypothetischer Verteilung) erwarteten (absoluten) Häufigkeiten, jeweils durch die

erwartete geteilt. Anschließend wieder derselbe Mechanismus: Zu großer Wert von χ^2 spricht gegen die Hypothese, dass die betrachtete Variable der vorgegebenen Verteilung folge. (Wieder die Asymmetrie: Man kann zu einem recht sicheren Verwerfen kommen, nicht aber bei Nichtverwerfen sicher sein, die Hypothese sei korrekt - allenfalls kann sie bei recht großer Stichprobe nicht mehr *allzu* falsch sein.) Dabei hat man lediglich noch die Zahl der Freiheitsgrade richtig zu bestimmen und damit zur entsprechenden χ^2- Verteilung zu gehen:

$df =$ Zahl der „Häuser" minus eins minus Zahl der geschätzten Verteilungsparameter.

Man beachte, dass auch diese Freiheitsgradbestimmung *abstrakt* derselben Regel folgt wie bei der Unabhängigkeitsfrage: Bei einer $(m \times n)-$ Tafel hat man $mn-1$ „Häuser", davon die Zahl der geschätzten Parameter - das waren in diesem Falle die Randwahrscheinlichkeiten - abgezogen, das ergibt $mn-1-(m-1)-(n-1) = mn-m-n+1 = (m-1)(n-1)$. Man beachte dabei, dass es nur $m-1$ unabhängige Zeilenwahrscheinlichkeiten (und entsprechend $n-1$ Spaltenwahrscheinlichkeiten) zu schätzen gibt, da deren Summe 1 jeweils ergeben muss. Das beschriebene Verfahren für die neue Problemstellung wollen wir an ein paar Beispielen illustrieren:

2.2. Beispiele. Erstes Beispiel: Zu prüfen ist die Hypothese, eine Variable X sei gleichverteilt auf dem Intervall $[0, 2]$. (Gleichverteilt heißt: Konstante Dichte, Histogrammbild mit nur einem Kasten. D.h. auf jeden Bereich in $[0, 2]$ gleicher Breite b entfällt dieselbe Wahrscheinlichkeit $b/2$. Besonderheit dieses Beispiels: Es ist eine ganz bestimmte Verteilung Gegenstand der Hypothese, daher sind keine Verteilungsparameter zu schätzen und später bei den Freiheitsgraden zu berücksichtigen. Nun möge man in einer Stichprobe von $X-$ Werten folgende Häufigkeiten beobachtet haben:

Wert	$0-0.4$	$0.4-0.8$	$0.8-1.2$	$1.2-1.6$	$1.6-2$
Anzahl	30	23	25	27	35

.

Es ist in diesem Falle sehr leicht, die Tafel der laut Hypothese zu erwartenden Häufigkeiten anzugeben: Gemäß Gleichverteilung erwarten wir für jedes Intervall $140 \cdot 0.4/2 = 28$. Offenbar weichen die beobachteten Häufigkeiten nur mäßig davon ab - doch schon zu viel? Rechnen wir aus:

$$\chi^2 = \frac{(30-28)^2}{28} + \frac{(23-28)^2}{28} + \frac{(25-28)^2}{28} + \frac{(27-28)^2}{28} + \frac{(35-28)^2}{28} = 3.143.$$

Das ergibt bei vier Freiheitsgraden ein Verwerfen der Hypothese auf dem (lächerlichen) Signifikanzniveau

$$\alpha = P\left(\chi^2_{df=4} \geq 3.143\right) > 0.5, \text{ (ziemlich genau 0.53).}$$

Damit ist die Wahrscheinlichkeit, die Hypothese fälschlich zu verwerfen, viel zu groß. Der Eindruck von einer einigermaßen guten Übereinstimmung war richtig, wenn wir auch keineswegs sicher sein können, dass eine größere Stichprobe nicht zum Verwerfen der Hypothese führen würde. Aber dann sollte man sich fragen, ob die statistisch „signifikanten" Abweichungen denn auch inhaltlich bedeutsam sind: Testen wir dieselbe Hypothese anhand des neuen Befundes:

Wert	$0-0.4$	$0.4-0.8$	$0.8-1.2$	$1.2-1.6$	$1.6-2$
Anzahl	300	230	250	270	350

.

Dann haben wir entsprechend (erwartet wären 280 in jedem Haus)

$$\chi^2 = 31.43.$$

Man mache sich klar, warum der zehnfache Wert für χ^2 resultiert und nicht etwa derselbe. Nun ist dieser Wert astronomisch groß für 4 Freiheitsgrade, man kann so gut wie sicher sein, dass die Hypothese falsch ist. Dennoch bleibt die Gleichverteilung ein einfaches Modell, das grob gesehen recht gut mit der tatsächlichen Verteilung übereinstimmt. Entsprechendes kann man mit großen Stichproben auch dann noch veranstalten, wenn ein hypothetisches Modell noch viel besser mit der tatsächlichen Verteilung einer Variablen übereinstimmt.

Zweites Beispiel: Bei einer Variablen X werden nur ganzzahlige Werte im Bereich von 0 bis 5 beobachtet. Zu testen ist die Hypothese, es handle sich um eine Binomialverteilung mit $n = 5$. In diesem Falle braucht man nicht zu gruppieren. Dafür entsteht die zusätzliche Komplikation gegenüber dem ersten Beispiel, dass ein Parameter (nämlich p) der Binomialverteilung empirisch zu schätzen verbleibt - erst dann können die Wahrscheinlichkeiten für die einzelnen Werte geschätzt und die erwarteten Häufigkeiten konkret angegeben werden. Beobachtet habe man:

Wert	0	1	2	3	4	5
Anzahl	65	190	135	75	10	5

.

Daraus lässt sich p einfach so schätzen: Bei 480 Beobachtungen gab es einen $X-$Mittelwert $\overline{x} = (0 \cdot 30 + 1 \cdot 230 + 2 \cdot 130 + ... + 5 \cdot 5)/480 = 780/480 = 13/8$, dies ist ein Schätzwert für den Mittelwert der hypothetischen Binomialverteilung, also für $n \cdot p$, was mit $n = 5$ genau $13/40$ als Schätzwert für p hervorbringt. (Unter den Binomialverteilungen mit $n = 5$ ist diejenige mit $p = 13/40$ am besten an die Daten angepasst.) Wir berechnen nunmehr die Tafel der erwarteten Häufigkeiten mittels der Wahrscheinlichkeitsformel für diese Binomialverteilung, zum Beispiel für 2 als $480 \cdot \binom{5}{2} (13/40)^2 (1 - 13/40)^{5-2}$, allgemein zum Wert i als $480 \cdot \binom{5}{i} (13/40)^i (1 - 13/40)^{5-i}$, $i = 0, ..., 5$:

Wert	0	1	2	3	4	5
Anzahl	67.261	161.92	155.93	75.076	18.074	1.7404

.

Damit wird

$$\begin{aligned} \chi^2_{df=4} &= \frac{(65-67.261)^2}{67.261} + \frac{(190-161.92)^2}{161.92} + \frac{(135-155.93)^2}{155.93} \\ &\quad + \frac{(75-75.076)^2}{75.076} + \frac{(10-18.074)^2}{18.074} + \frac{(5-1.7404)^2}{1.7404} \\ &= 17.467. \end{aligned}$$

Das ist entschieden zu viel, um noch an eine zugrundeliegende Binomialverteilung glauben zu können: Auf einem Niveau $\alpha < 0.005$ (ziemlich genau 0.00157) kann die Hypothese verworfen werden. Die Häufigkeit des Wertes 1 steht zu weit heraus, der Abfall beim Wert 4 ist zu drastisch. Das sind charakteristische qualitative Abweichungen von einer Binomialverteilung. (Dergleichen kann man übrigens durchaus praktisch benutzen: So haben wir einmal bei einer (Spiegel-) Umfrage unter Studenten herausbekommen, dass einige Daten aller Wahrscheinlichkeit nach gefälscht

waren: Es gab diskrete Notenwerte, deren Verteilungen fast stets sehr exakt Binomialverteilungen waren, in seltensten Ausnahmefällen (die auch sonst etwas unglaubwürdig waren) jedoch drastisch davon abwichen und bei weitem keinen solchen χ^2- Test bestanden.

Häufig wird man Daten daraufhin überprüfen, ob sie einer Normalverteilung entstammen. Das ist insbesondere dann wichtig, wenn man statistische Verfahren benutzt, deren Gültigkeit Normalverteilungen voraussetzen und die empfindlich reagieren auf drastische Verletzungen dieser Voraussetzung. Dabei treten beide an den Beispielen besprochenen Komponenten auf: Man hat zu gruppieren und außerdem zwei Parameter, μ und σ, zu schätzen anhand der Daten, natürlich durch $\overline{x}$ und $s(X)$. Anschließend benutzt man die χ^2- Verteilung mit den Freiheitsgraden: Anzahl der Häuser minus 1 minus 2. Bei Gruppierung sollte man stets ein vernünftiges Maß der Klassenzahl finden: Zu grobe Klassifikation würde zu wenige Formelemente der Verteilung überprüfen, mit der Gefahr, dass eine recht falsche Verteilungshypothese nicht zu verwerfen wäre. (Zum Beispiel würde *jede* symmetrisch verteilte Variable jeden χ^2- Test auf Normalverteiltheit bestehen, wenn man in nur zwei Intervalle einteilen würde, deren Grenze bei der Symmetriemitte liegt.) Andererseits führt zu feine Einteilung zu einer Unterbesetzung einzelner Felder, wie sie für einen χ^2- Test ungünstig ist (Faustregel: Mindestens 5 in jedem Haus).

KAPITEL 5

Regression und Korrelation

1. Exakte funktionsartige Zusammenhänge zwischen Variablen, linear und nichtlinear

Eine sprachliche Vorbemerkung ist notwendig: Der Begriff der Unabhängigkeit tritt in der Mathematik auf vielfältigste Weise auf, und bisher verwandten wir ihn bereits häufig in zwei *völlig verschiedenen Bedeutungen:* Variablen X, Y (auch mehr noch als zwei) heißen (*im statistischen Sinne*) unabhängig, wenn stets gilt, dass $P(X \leq a \text{ und } Y \leq b \text{ und...}) = P(X \leq a) \cdot P(Y \leq b) \cdot ....$ Dagegen abzugrenzen ist der Sprachgebrauch, dass bei Funktionen die unabhängige Variable für ein beliebiges Element des Definitionsbereiches steht und die abhängige Variable der davon exakt abhängige Funktionswert ist. Nun ist es eine Besonderheit gerade dieses Kapitels, dass *beide Bedeutungen* ständig nebeneinander vorkommen. Das liegt genau darin begründet, dass hier ins Auge gefasst wird, eine Variable X im Sinne der Statistik als unabhängige Variable und eine andere statistische Variable Y als abhängige Variable, also $Y = f(X)$ aufzufassen, mindestens näherungsweise. Daher werden wir stets ausführlicher die statistische Unabhängigkeit mit dem Adjektiv „statistisch" versehen. Steht bloß „unabhängig", so ist das einfach im Sinne von „unabhängige Variable" gemeint. Ferner tritt besonders auch die „lineare" Unabhängigkeit von Variablen auf, was stets die Abschwächung der statistischen Unabhängigkeit bedeutet (bei zwei Variablen X, Y: $Cov(X, Y) = 0$).

Wir setzen stets voraus, dass wir so etwas wie einen funktionalen Zusammenhang zwischen zwei Variablen X und Y betrachten, die *denselben* Definitionsbereich Ω haben. Wir haben gesehen, dass solche Variablen *statistisch unabhängig* sein können, dann ergibt die Kenntnis des X-Wertes keinerlei Information über den Y-Wert. Antipodisch dazu steht der Fall, dass der Y-Wert durch den X-Wert *eindeutig* bestimmt ist; das heißt: Es gibt eine Funktion f, so dass stets gilt: Y-Wert $= f(X\text{-Wert})$. Wir können damit X als unabhängige Variable, Y als abhängige Variable auffassen. (Dies ist der einfachste Fall, in komplexeren Situationen wird man stets mehrere unabhängige und auch abhängige haben.) Also genau: Für alle $\omega \in \Omega$ gilt: $Y(\omega) = f(X(\omega))$. Kürzer formuliert man das gern so: $Y = f(X)$, exakter müsste es „$Y = f$ hinter X geschaltet" heißen. f ist dabei eine Funktion $\mathbb{R} \to \mathbb{R}$, und sie rechnet die Y-Werte aus den X-Werten aus, man kann sie nicht etwa auf eine Variable wie X *anwenden.* Man beachte weiter, dass eine solche funktionale Abhängigkeit eine *Richtung* hat; gibt es eine in der einen Richtung, so gibt es nicht zwangsläufig auch eine für die andere Richtung. Das hängt daran, ob die Funktion im relevanten Gebiet umkehrbar ist.

Beispiele für exakte funktionale Abhängigkeit:

$X =$ Preis in DM, $Y =$ Preis in Dollar (bei einem festgelegten Wechselkurs). In diesem Falle gilt $Y = c \cdot X$ [detaillierter gesagt noch einmal: Für jede Ware ω gilt: $Y(\omega) = f(X(\omega))$], der Zusammenhang wird vermittelt durch eine lineare

Funktion, sogar Proportionalität. Mit dem X-Wert liegt der Y-Wert eindeutig fest. Ein ähnliches Beispiel: X = Temperatur in Celsius, Y = Temperatur in Fahrenheit. Wieder hat man eine lineare Funktion, allerdings mit einer additiven Konstanten.

Ein nichtlinearer (aber auch exakter funktionaler) Zusammenhang: Y = zurückgelegter Weg bei einem freien Fall, X = abgelaufene Zeit nach „Loslassen". Hier hat man $Y = \frac{1}{2}X^2$. Das ist noch extrem einfach, es gibt wichtige wesentlich komplexere funktionale Zusammenhänge.

Wir wollen in der Statistik jedoch auf *inexakte* (aber dennoch „ungefähr" funktionale) Zusammenhänge hinaus, und so etwas besprechen wir im Prinzip zu Beginn des Abschnittes 2., im Einzelnen für den einfachsten linearen Fall in 2.1, verallgemeinert in 2.2. Zum Verständnis ist es jedoch wichtig, erst einmal zu wissen, wie ein exakter funktionaler Zusammenhang aussieht. Das haben wir für den Fall einer unabhängigen und abhängigen Variablen besprochen. Wenigstens ein Beispiel wollen wir noch dafür bringen, dass man zwei unabhängige Variablen hat:

Population: Alle rechtwinkligen Dreiecke, X_1 = Länge der ersten Kathete (unwichtig, wie man das festlegt, für unser Beispiel ist das auch gleichgültig), X_2 = Länge der zweiten Kathete, Y = Länge der Hypotenuse. Dann können wir nach dem Satz des Pythagoras Y als Funktion von X_1 und X_2 folgendermaßen ausrechnen: $Y = X_1^2 + X_2^2$. Die Hypotenusenlänge (als abhängige Variable) ist also eine nichtlineare Funktion beider Kathetenlängen (als unabhängiger Variablen). Ein solcher Zusammenhang ist stets theoretischer Art, und man nennt so etwas daher auch „Modell". Das Modell des funktionalen Zusammenhangs lässt sich in einer Gleichung ausdrücken, und daher findet man zuweilen den etwas irreführenden und unglücklichen Ausdruck, diese Gleichung sei das Modell. Man lasse sich aber nicht verleiten, zu glauben, diese simplen Modelle seien alles, was die Mathematik an Modellen zu bieten hätte! Aber selbst beim funktionalen Zusammenhang handelt es sich um ein „Modell" durchaus in einem tieferen Sinne: Es wird mit einer mathematischen Rechenoperation aus den Werten der Einflussvariablen ein Wert produziert, den etwa die Natur unter den Bedingungen, welche durch die Werte der Einflussvariablen beschrieben werden, auf eine ganz andere Weise produziert, mit unendlichen Komplikationen. Es ist eine gewaltige metaphorische Übertragung, wenn man sagt, die Natur „rechne".

Wir fassen noch einmal zusammen:

Gleichung für das Modell eines funktionalen Zusammenhangs zwischen einer abhängigen Variablen Y und unabhängigen Variablen $\mathbf{X_1}, ..., \mathbf{X_n}$:

$$\mathbf{Y = f(X_1, ..., X_n)} \tag{1.1}$$

Man beachte: Es ist nichts über die Art von f gesagt, diese Funktion könnte beliebig einfach oder kompliziert sein, und sie ist stets nur im konkreten Einzelfall gegebener Variablen zu spezifizieren, beispielsweise lautet das *lineare* Modell mit nur einer unabhängigen Variablen: $Y = aX + b$.

In jedem Einzelfall stellen sich folgende Grundfragen: Gibt es überhaupt einen derartigen Zusammenhang zwischen den interessierenden Variablen, und wenn ja, welcher Art ist er dann, welche Funktion f beschreibt diesen Zusammenhang korrekt (oder doch wenigstens: genau genug - dann kommen mehrere verschiedene Funktionen in Frage, die sich durchaus zuweilen nach rationalen Prinzipien auswählen lassen)? (Diese Grundfragen werden sich bei Betrachtung *inexakter* funktionaler Zusammenhänge ein wenig modifizieren.) Diese Grundfragen sind je nach

theoretischem oder empirischem Zusammenhang in völlig verschiedenartigen Weisen anzugehen, das kann mathematisch-theoretisch deduzierend sein, aber in empirischer Wissenschaft auch stark anhand gegebenen Datenmaterials vorgehen (dazu mehr im nächsten Abschnitt).

Abschließend bemerken wir noch, dass man im Zusammenhang mit solchen Modellen gern die unabhängigen Variablen auch „Prädiktorvariablen" oder einfach „Prädiktoren" nennt, das heißt „voraussagende Variablen" - man denkt dabei an den Zweck, den Wert einer interessierenden „Zielvariablen" (so nennt man dann gern die abhängige Variable) *vorauszusagen*, wenn man die Werte der Prädiktorvariablen kennt. Beispiel: Man möchte aus sozialwissenschaftlichen Daten über die soziale Umgebung eines Kindes dessen späteren Berufserfolg etc. voraussagen. So etwas geht natürlich nicht exakt, aber immerhin kann man einige Variablen als wesentliche EinflussVariablen erkennen und auch genauer erweisen. (Dieser Aspekt wird im nächsten Abschnitt vertieft.)

2. Inexakte (statistische) funktionsartige Zusammenhänge

Es sollte naheliegen, die Inexaktheit einfach dadurch ins Spiel zu bringen, dass man aus der Modellgleichung 1.1 eine Ungefähr-Gleichung macht, also formuliert:

$$\mathbf{Y \approx f(X_1, ..., X_n)}.$$

Dies trifft jedoch nicht stets das Gewünschte, einerseits möchte man spezifizieren, wie gut denn das „ungefähr" ist (auf manchen Gebieten für mancherlei Zwecke verlangt man sehr kleine Abweichungen), welche Fehler man also zu erwarten hat, andererseits möchte man zumal auf Gebieten, in denen die meisten Einflussvariablen auf eine Zielvariable unbekannt sind und bleiben, z.B. auf dem der Psychologie oder Sozialwissenschaft, vernünftigerweise nicht gleich darauf hinaus, einen Wert aus Werten von Einflussvariablen gleich mit guter Genauigkeit zu reproduzieren, sondern man gibt sich durchaus mit einem großen Fehler zufrieden („ungefähr" ist dann viel zu viel gesagt). Aber man möchte dann wenigstens wissen und beschreiben, dass die betrachteten Einflussvariablen den Wert der Zielvariablen in *nennenswertem Maße* erklären, also einen guten Teil davon. ein wenig bemerkenswert, aber durchaus typisch: Die Mathematik stellt ein Mittel bereit, das gleichermaßen geeignet für beide so verschieden erscheinede Zwecke ist. Der Kunstgriff ist sehr einfach und besteht darin, aus der „ungefähr"-Gleichung eine genaue Gleichung zu machen und das nicht Aufgehende gesondert als damit implizit definierte Fehlervariable oder Rest aufzuführen (statistisches Gegenstück zur Formel 1.1):

$$\mathbf{Y = f(X_1, ..., X_n)+E}. \tag{2.1}$$

(E für „error", d.h. Fehler.) E ist also *definitionsgemäß* die Differenz $Y - f(X_1, ..., X_n)$. Die Fehlervariable zu beschreiben, das heißt die Qualität des inexakten funktionalen Zusammenhangs zu beschreiben. Der exakte Fall ist gleich mit enthalten: Dabei ist einfach die Fehlervariable die Konstante mit Wert Null. Wir wollen nunmehr sehen, wie man sowohl für hohe Genauigkeit als auch für geringe etwas Substantielles über E aussagen kann. Im ersteren Fall würde man etwa als Resultat befriedigend finden, der Wert von E betrage stets (oder mit spezifizierter

hoher Wahrscheinlichkeit) nur ein Prozent (oder gar ein Tausendstel etc.) vom Sollwert, dem Y-Wert. Im letzteren Falle, z.B. in der Sozialwissenschaft oder Psychologie, formuliert man typisch Resultate der Form: „Der Anteil der Varianz der Zielvariablen Y, der durch den funktionalen Zusammenhang erkärt wird (also durch den Wert $f(X_1, ..., X_n)$), beträgt ...". Ebenso gut kann man das ausdrücken durch den Anteil, den die Varianz von E an der Varianz von Y hat, also durch den Quotienten $\sigma^2(E)/\sigma^2(Y)$. (Genauer ist dies dann der Fall, wenn die Variablen $f(X_1, ...X_n)$ und E linear unabhängig sind, so dass gilt: $\sigma^2(Y) = \sigma^2(f(X_1, ..., X_n)) + \sigma^2(E)$.) Somit lohnt es in beiden Fällen, die Eigenschaften von E zu beschreiben, um die Qualität der versuchten funktionalen Erklärung zu erfassen. Nun stellen wir das Grundproblem konkreter für den Fall einer empirischen Wissenschaft, die in der Situation steht, dass man keine tieferen theoretischen Mittel zur Hand hat, sondern sich stark an gegebene empirischen Daten halten muss:

Grundsituation und Grundproblem:

Man hat empirische Daten der Form $(x_1^{(i)}, ...x_n^{(i)}, y^{(i)})$, $1 \leq i \leq n$, d.h. n Beispiele für Werte der Prädiktoren und den zugehörigen Wert der Zielvariablen. Gesucht ist eine Funktion f, die für die gegebenen Beispiele (die gegebene Stichprobe) den besten funktionalen Zusammenhang darstellt. Dies Problem ist durchaus nicht ganz einfach, vielmehr ist man zunächst darauf angewiesen, aus der Durchsicht der Daten eine günstige Klasse möglicher Funktionen, gewöhnlich mit Parametern definiert, auszusuchen und auszuprobieren. Niemals weiß man (sofern man nicht über tiefere theoretische Prinzipien verfügt), ob es noch bessere gibt oder nicht. Allerdings verhilft die Mathematik dann zu zweierlei: Erstens kann man mit mathematischer Extremwertrechnung die optimale Funktion aus der anvisierten Klasse aussondern (tatächlich ausrechnen), zweitens dann beurteilen, wie weit man damit gekommen ist. Reicht das Resultat für die gegebenen Zwecke aus, so ist es gut, andernfalls hat man immerhin die Möglichkeit, sich eine günstigere Klasse von Modellfunktionen auszudenken. Man kommt also auch zu einem vernünftigen Urteil darüber, was die ausprobierte Klasse taugt.

Nunmehr wenden wir uns der Lösung des Optimalitätsproblems speziell für eine *lineare* Funktion f zu und beschränken uns zunächst auf eine einzige unabhängige Variable. (Für Verallgemeinerungen vgl. Abschnitt 2.2.)

2.1. Linearer Zusammenhang zwischen zwei Variablen: Regressionsgerade. Wir betrachten hier speziell das lineare Modell (also mit linearer Funktion f) für nur eine einzige Prädiktorvariable, die zugehörige Gleichung lautet:

$$Y = aX + b + E. \tag{2.2}$$

Wir setzen in den folgenden Abschnitten *generell voraus*, dass $\sigma^2(Y) \neq 0$ und $\sigma^2(X) \neq 0$. Wäre nämlich $\sigma^2(Y) = 0$, so wäre Y eine Konstante, und wir können Y perfekt durch diese Konstante b mit $a = 0$ wiedergeben, das ganze Problem wäre ins völlig Uninteressante trivialisiert. Wäre dagegen $\sigma^2(X) = 0$, so wäre X eine Konstante, ohne Varation, und der Wert von X könnte keinerlei Information über variierende Werte von Y geben, es wäre also von vornherein unmöglich, Y auch nur zu einem kleinen Teil als Funktion von X vorauszusagen, unser Projekt wäre von vornherein gescheitert.

Es seien X und Y nun zwei beliebige Variablen. Wie sollte man dann die Parameter a und b optimal wählen, so dass von der Variation der Y-Werte so

viel wie möglich durch die lineare Funktion von X erklärt wird, der Fehler E also minimalisiert? Wie wäre „Optimum" präziser zu fassen? Ein sehr einfacher Ansatz dazu besteht darin, folgende geringfügigen und selbstverständlichen Forderungen an die Wahl von a und b zu stellen:

$$\begin{aligned} 1)\ \mu(E) &= 0 \\ 2)\ Cov(aX+b,E) &= 0. \text{ (Für } a \neq 0 \text{ gleichwertig: } Cov(X,Y)=0.) \end{aligned}$$

Erinnerung: $Cov(X,Y) = \mu[(X-\mu(X))\cdot(Y-\mu(Y))]$.

Zur Selbstverständlichkeit: Wäre $\mu(E) \neq 0$, so könnte man sofort den Mittelwert $\mu(E)$ der Konstanten b zuschlagen und hätte dann eine neue Fehlervariable mit Mittelwert 0. Die zweite Forderung ist deswegen vernünftig, weil eben der ganze lineare Zusammenhang zwischen X und Y in $aX+b$ enthalten sein sollte, eine weitere lineare Abhängigkeit zwischen E und $aX+b$ (gleichwertig zwischen E und X, wenn nur $a \neq 0$) also nicht mehr bestehen sollte. Allerdings werden wir erst mit dem hier herzuleitenden Resultat diese Bedeutung der Forderung 2) an die Kovarianz erkennen. Dagegen wissen wir bereits, dass 2) die Bedeutung einer Abschwächung der *statistischen* Unabhängigkeit zwischen E und X besitzt. Erstaunlich ist es, dass diese schwachen Forderungen genügen, um bereits a und b eindeutig zu bestimmen! Später werden wir auch sehen, in welchem Sinne damit die Fehler minimalisiert werden und die mittels 1), 2) ausgesonderte Lösung optimal ist.

Wir folgern nunmehr aus 1) und 2), wie a und b aussehen müssen:

Kovarianz Null bedeutet, dass Summenbildung und Varianz verträglich sind (vgl. den Abschnitt über das Rechnen mit Mittelwerten und Varianzen, den wir auch für weitere Rechnungen hier stark benutzen müssen), also haben wir:

$$\sigma^2(Y) = \sigma^2(aX+b) + \sigma^2(E) = \sigma^2(aX+b) + \sigma^2(Y-aX-b). \text{ (Def. von } E!)$$

Nun gilt ganz allgemein, dass $\sigma^2(X) = \mu((X-\mu(X))^2) = \mu(X^2) - \mu^2(X)$ (dabei ist $\mu^2(X) = (\mu(X))^2$), auch hat man das Analogon dazu für die Kovarianz: $Cov(X,Y) = \mu(XY) - \mu(X)\mu(Y)$, also erhalten wir unter Benutzung der Tatsache, dass Addition einer Konstanten zu einer Variablen nichts an der Varianz ändert:

$$\begin{aligned} \mu(Y^2) - \mu^2(Y) &= a^2\sigma^2(X) + \mu((Y-aX)^2) - \mu^2(Y-aX) \\ &= a^2\mu(X^2) - a^2\mu^2(X) + \mu(Y^2) - 2a\mu(XY) + a^2\mu(X^2) \\ &\quad -\mu^2(Y) + 2a\mu(X)\mu(Y) - a^2\mu^2(X) \\ &= \mu(Y^2) - \mu^2(Y) \\ &\quad -2a\mu(XY) + 2a\mu(X)\mu(Y) + 2a^2\mu(X^2) - 2a^2\mu^2(X) \\ &= \mu(Y^2) - \mu^2(Y) - 2aCov(X,Y) + 2a^2\sigma^2(X). \end{aligned}$$

Damit resultiert die Gleichung

$$a = \frac{Cov(X,Y)}{\sigma^2(X)}. \text{ (Wir setzten } \sigma^2(X) \neq 0 \text{ voraus!)}$$

Außerdem ergibt die erste Forderung sofort

$$\mu(E) = \mu(Y-aX-b) = \mu(Y) - a\mu(X) - b = 0,$$

also

$$b = \mu(Y) - a\mu(X).$$

Bemerkung: Eleganter wird die Rechnung, wenn man sie unter der Voraussetzung $\mu(X) = \mu(Y) = 0$ ausführt und von da aus theoretisch bequem auf den allgemeinen Fall schließt, was dem Anfänger jedoch eine zusätzliche Schwierigkeit bedeuten kann.
Damit haben wir die Parameter der (besten, s.u. Abschnitt 2.1.3) linearen Funktion zur Voraussage der Y-Werte aus den X-Werten bestimmt. Man nennt die zugehörige Gerade **Regressionsgerade**, und die soeben berechneten Parameter heißen **Regressionsparameter**. Um die Richtung der Voraussage mit zu bezeichnen, notiert man auch gern in unseren Rechenergebnissen (der Punkt dient hier nur zur Trennung und kann auch fehlen, man denke hier nicht an „mal"!):

$$a_{Y\cdot X} = \frac{Cov(X,Y)}{\sigma^2(X)},\ b_{Y\cdot X} = \mu(Y) - a\mu(X). \tag{2.3}$$

Man beachte, dass für die andere Voraussagerichtung (von Y auf X) *andere* Werte $a_{X\cdot Y}$, $b_{X\cdot Y}$ herauskommen, natürlich mit den analogen Formeln, die durch Vertauschen der Buchstaben X, Y entstehen.

Eine erste Bemerkung zur Interpretation des Resultats für a: Wenn $a = 0$ herauskommt, so bedeutet dass: Die Variation von X trägt linear nichts zur Erklärung der Variation von Y bei, d.h. X und Y sind linear unabhängig. Aber $a = 0$ bedeutet nach unserer Formel, dass $Cov(X,Y) = 0$. Damit haben wir erklärt, warum diese Kovarianzbedingung mit Recht „lineare Unabhängigkeit" genannt wird.

2.1.1. *Der Korrelationskoeffizient $\rho(X,Y)$ als Maß für die Stärke des linearen Zusammenhangs zwischen X und Y.* Wir schauen nach der Qualität der linearen Voraussage der Y-Werte aus den X-Werten, d.h. danach, wie hoch die Varianz des Fehlers, $\sigma^2(E)$, relativ zur Varianz der Zielvariablen Y ist. Wieder ist unsere Varianzzerlegung gemäß Forderung 2) entscheidend:

$$\sigma^2(Y) = \sigma^2(aX + b) + \sigma^2(E) = a^2\sigma^2(X) + \sigma^2(E).$$

Wir dividieren diese Gleichung durch $\sigma^2(Y)$ (das ist nach Voraussetzung $\neq 0$) und erhalten, indem wir den ausgerechneten Wert für a einsetzen:

$$1 = \frac{Cov^2(X,Y)\sigma^2(X)}{\sigma^2(Y)\sigma^4(X)} + \frac{\sigma^2(E)}{\sigma^2(Y)} = \frac{Cov^2(X,Y)}{\sigma^2(Y)\sigma^2(X)} + \frac{\sigma^2(E)}{\sigma^2(Y)}.$$

Hier tritt im ersten Summanden eine interessante Größe auf, durch die man offenbar den Anteil der Fehlervarianz an der Varanz von Y ausdrücken kann:

Definition 17 (Korrelationskoeffizient zweier Variablen). *Der Korrelationskoeffizient der Variablen X, Y mit nichtverschwindenden Streuungen, bezeichnet mit $\rho(X,Y)$, ist definiert als*

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{2.4}$$

Damit haben wir folgende Zusammenhänge: $\rho^2(X,Y)$ (der erste Summand oben) ist der Anteil der Varianz der Zielvariablen Y, welcher durch lineare Regression auf X erklärt ist, und entsprechend ist

$$1 - \rho^2(X,Y) = \frac{\sigma^2(E)}{\sigma^2(Y)}. \tag{2.5}$$

der Anteil der Varianz von Y, der *nicht* durch den *linearen* Zusammenhang aus X erklärt wird.

Folgerung: Es gilt stets

$$-1 \leq \rho(X,Y) \leq 1.$$

Denn in der vorigen Formel steht auf der rechten Seite eine Zahl ≥ 0, somit muss $\rho^2(X,Y) \leq 1$ gelten, also kann der Korrelationskoeffizient nur Werte im Bereich $[-1,1]$ annehmen.

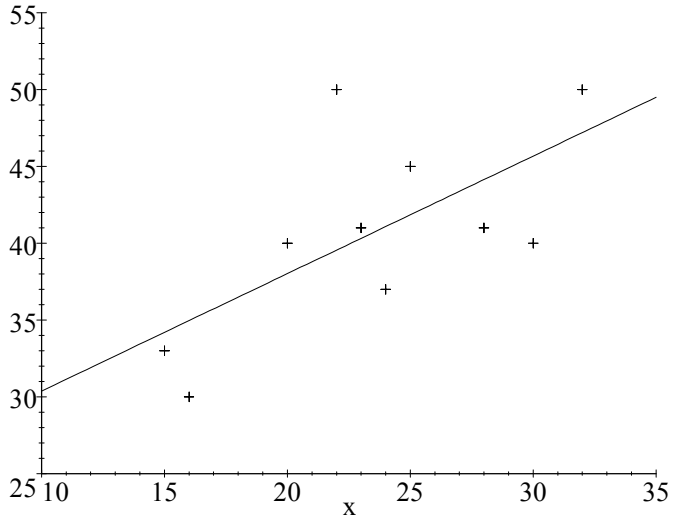
Übrigens kann man (mit kleiner Ungenauigkeit) $\sigma(E)$ als mittleren Fehler der Voraussage der Y-Werte durch die X-Werte interpretieren.

Beispiel:

Wir nehmen eine Gruppe von 10 Studenten als Gesamtpopulation (wir reden also von μ, ρ, a, b statt von Schätzwerten $\overline{x}, \overline{y}, r, \widehat{a}, \widehat{b}$ - vgl. zu diesen den nächsten Abschnitt 2.1.2), die Variable X sei die Punktezahl in einer Statistikklausur, die Variable Y die Punktezahl in einer Methodenklausur. Die zugehörige Wertetabelle aus den Wertepaaren (X-Wert, Y-Wert) („bivariat" nennt man das in gewissen Kreisen gern, „multivariat", wenn es noch mehr als zwei sind, also bei mehreren Prädiktoren) sehe so aus:

$(20,40), (25,45), (16,30), (30,40), (22,50), (15,33), (28,41), (32,50), (23,41), (24,37)$.

Zunächst sollte man den zugehörigen Punkteschwarm in der (X,Y)-Ebene aufzeichnen, das ergibt die Kreuze in folgender Graphik:



Man sieht im Beispiel, dass es Leute gab, die bei der einen Klausur relativ gut und bei der anderen nicht so gut waren, aber immerhin ist ein linearer Trend deutlich erkennbar (in vergleichbaren Beispielen kommt übrigens meist ein noch

viel besserer linearer Zusammenhang als hier heraus). Wir können folgende Werte ausrechnen - das geht auch schon bequem mit jedem Taschenrechner, der Paardaten aufzunehmen fähig ist):

$$\rho = 0.65692.$$

(Vielfach kann man 0.8 und mehr bei solchen Beispielen sehen.) Das ist immerhin etwas. Der Anteil der Varianz von Y, der durch lineare Regression auf X erklärt wird, ist also $\rho^2 = 0.43154$ oder etwa 43%. Die Regressionsparameter kann man ebenfalls auf einem Taschenrechner ablesen, wenn auch nicht in dieser genauen Form von Brüchen:

$$\begin{aligned} a &= \frac{13}{17}, \\ b &= \frac{1932}{85}, \text{ die Regressionsgerade lautet also} \\ Y &= \frac{13}{17}X + \frac{1932}{85}. \end{aligned}$$

Sie wurde in der Graphik oben bereits eingezeichnet. (Rechnung sowie Zeichnung anzufertigen ist äußerst langweilig und sollte man am besten einem geeigneten Computerprogramm überlassen, man muss dazu lediglich lernen, in welcher Form die Daten einzugeben sind, das ist bei verschiedenen Programmen immer wieder anders und natürlich erst recht äußerst langweilig.)

Bemerkung 1: Dieselben Werte erhält man als Schätzwerte für die richtigen Werte, wenn es sich nur um eine Stichprobe handelt, allerdings müssen die Stichproben recht groß sein, wenn sie gut, also mit einiger Sicherheit einigermaßen genau sein sollen, vgl. dazu den nächsten Abschnitt 2.1.2.

Bemerkung 2: Wir wollen hier einmal quantitativ nachschauen, wie genau die Voraussage der Y-Werte durch die X-Werte in unserem Beispiel ist. (Wir wissen natürlich bereits, dass die Regressionsgerade die beste lineare Vorhersage gibt, und wir sehen im Beispiel am Punkteschwarm keinen Anlass, es mit einer nichtlinearen Funktion zu versuchen.) Dazu betrachten wir den mittleren quadratischen Fehler der Vorhersage, das ist einfach

$$\frac{1}{10}\sum_{i=1}^{10}\left(y_i - \frac{13}{17}x_i - \frac{1932}{85}\right)^2 = \sigma^2(E) = \sigma^2(Y)\cdot(1-\rho^2).$$

Das brauchen wir also gar nicht konkret auszurechnen, vielmehr lesen wir $\sigma^2(Y)$ sofort aus dem Rechner ab, und ρ^2 haben wir bereits. Es ergibt sich in unserem Beispiel der mittlere quadratische Fehler der linearen Vorhersage

$$38.01\cdot(1-0.65692^2) = 21.607.$$

Mit leichter Ungenauigkeit können wir die Wurzel davon, das ist natürlich $\sigma(E)$, als mittleren absoluten Fehler dieser Vorhersage angeben, also etwa 4.65. Das bedeutet also, dass wir im Mittel die Punktezahl der zweiten Klausur mit nur etwa 4.65 Punkten Fehler, das ist etwa 10% von der Größenordnung der vorauszusagenden Werte, voraussagen können, und das ist recht gut, es macht keinen Notenunterschied. Im *Einzelfall* liegt die Voraussage natürlich auch einmal gründlich daneben! Der mittlere absolute Fehler ist natürlich *genau genommen* nicht dasselbe wie die Wurzel aus dem mittleren quadratischen Fehler, im Beispiel wollen einmal die Werte vergleichen:

Der mittlere absolute Voraussagefehler im Beispiel beträgt

$$\frac{1}{10}\sum_{i=1}^{10}\left|y_i - \frac{13}{17}x_i - \frac{1932}{85}\right| \approx 3.812.$$

Wir sehen also, dass der Unterschied zur Wurzel aus dem mittleren quadratischen Fehler nicht allzu groß ist. Dafür haben wir mit dem quadratischen Fehler (der Varianz) eine Größe, mit der sich systematische theoretische Rechnungen wesentlich leichter und übersichtlicher gestalten als solche mit Absolutwerten - alle allgemeinen Resultate insbesondere dieser Abschnitte wären damit nicht annähernd möglich!Schätzwerte für die Regressionsparameter und den Korrelationskoeffizienten anhand von Stichproben

Alle interessierenden Zahlwerte berechnen sich durch Kovarianz und Varianzen, und so liegt es nahe, die übliche Varianzschätzung anhand von Stichproben einzusetzen, die man analog auch für Kovarianzen durchführt:

$$\begin{aligned} s^2(X) &= \widehat{\sigma^2(X)} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}, \\ \widehat{Cov(X,Y)} &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}. \end{aligned}$$

Damit erhalten wir (man beachte, dass sich die Nenner $n-1$ wegkürzen):

$$r(X,Y) = \widehat{\rho(X,Y)} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}, \tag{2.6}$$

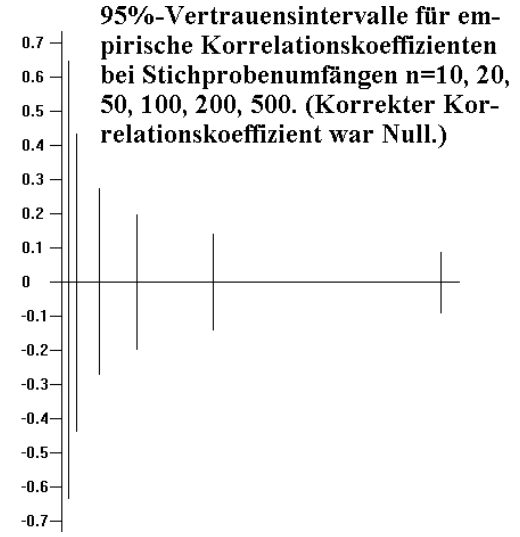
$$\widehat{a_{Y\cdot X}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

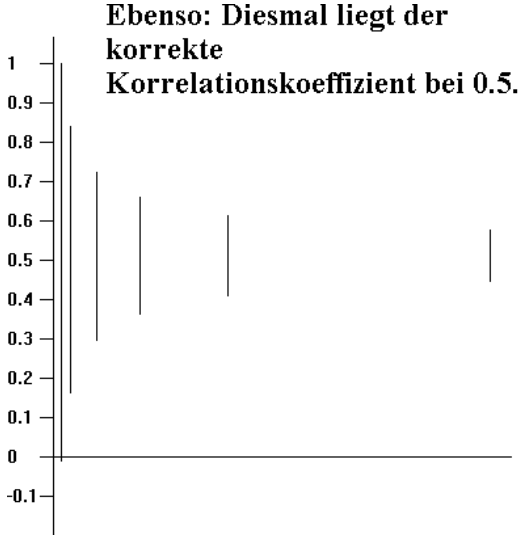
Man nennt diese auch den „empirischen" Korrelationskoeffizienten und die „empirische" Regressionssteigung. Bemerkung: Wenn die x_i, $1 \le i \le n$, alle Werte der Gesamtpopulation sind, zu den einzelnen Populationsmitgliedern, mit eventuellen Wiederholungen also, dann kommen bei diesen Formeln natürlich ρ und a heraus!

Wie genau ist die Schätzung eines Korrelationskoeffizienten (und entsprechend der Regressionssteigung) anhand einer Stichprobe?

Folgende Graphiken zeigen auf, dass man für vernünftige Qualität unangenehm hohe Stichprobenumfänge benötigt: Für verschiedene Stichprobenumfänge zeigen sie die Verteilung der empririschen Korrelationskoeffizienten und insbesondere die Breite eines 95%-Vertrauensintervalls für den Korrelationskoeffizienten. Dabei haben wir die idealtypische Situation zugrundegelegt, dass in Wirklichkeit gilt: $Y = X + E$, mit einer normalverteilten Variablen X und einer von X *statistisch* unabhängigen und ebenfalls normalverteilten Variablen E, deren Mittelwert Null ist. (Man vermute den Grund für die halbwegs deprimierenden Ergebnisse

also nicht etwa darin, dass die betrachteten Variablen überhaupt das Korrelationskonzept nicht korrekt anwenden lassen, im Gegenteil handelt es sich um den mathematisch idealen Fall der Anwendbarkeit, der so genau für empirische Variablen niemals vorliegt.) Den richtigen Korrelationskoeffizienten ρ kennen wir natürlich, wir können ihn über $\sigma^2(E)/\sigma^2(Y) = 1 - \rho^2$ beliebig einstellen. Wir wählen im Beispiel den Wert $\rho = 0$ und im zweiten den Wert $\rho = 0.5$ (das ist in *manchen* Zusammenhängen (Eignungstests etc.) schon hoch).





2.1.2. *Die Regressionsgerade als Lösung eines Extremwertproblems.* Hier leiten wir die Regressionsgerade noch einmal her als „optimale" Gerade zu einem Punkteschwarm (x_i, y_i), $1 \leq i \leq n$, dabei lernt man die wichtige und verbreitete „Kleinste-Quadrate-Methode" (engl. 'least squares') kennen. Für eine Gerade mit (beliebigen) Parametern a, b hat man den Voraussagefehler $y_i - ax_i - b$ für das i-te

Beobachtungspaar. Damit ist

$$f(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2 = \text{ quadratischer Gesamtfehler bei Vorhersage von } Y \text{ durch die Variable } aX+b$$

ein vernünftiges Maß für die Qualität der Vorhersage. Die beste Gerade wäre demnach gerade die, für welche $f(a,b)$ minimal wird. (Wir werden gleich sehen, dass es eine einzige Gerade gibt mit minimalem quadratischem Gesamtfehler.) Dies ist eine sehr einfache Extremwertaufgabe, ein wenig erschwerdend ist es nur, dass die Funktion f, deren Wert minimalisiert werden soll, eine Funktion von zwei unabhängigen Variablen ist. Für den vorliegenden konkreten Fall können wir dies Problem umgehen, stellen seine Lösung jedoch im nächsten Abschnitt 2.2 vor, der sich dem verallgemeinerten Regressionsproblem widmet.

Lösung des Extremwertproblems:

Zunächst setzen wir mit derselben Begründung wie oben wieder $b = \mu(Y) - a\mu(X)$, was wir hier konkreter $\overline{y} - a\overline{x}$ schreiben können, da wir ausschließlich die vorgegebenen Wertepaare betrachten wollen. Dann reduziert sich unser Problem auf die Aufgabe,

$$g(a) = \sum_{i=1}^{n}(y_i - ax_i - \overline{y} + a\overline{x})^2$$

zu minimalisieren. Wir bilden die erste Ableitung und setzen sie gleich Null:

$$g'(a) = 2\sum_{i=1}^{n}(\overline{x} - x_i)(y_i - ax_i - \overline{y} + a\overline{x}) = 0.$$

Das ist eine simple lineare Gleichung für a, man hat lediglich die Glieder mit dem Faktor a zu isolieren, auszuklammern und kann auflösen (fleißig $\sum x_i = n\overline{x}$ benutzen!) zu:

$$a = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}.$$

Die letzte Gleichung ergibt sich daraus, dass man wieder $\sigma^2(X) = \mu(X^2) - \mu^2(X)$ und das Analogon für die Kovarianz benutzt, was man leicht konkret ausrechnet. Das Resultat ist dasselbe wie oben. Man sieht leicht, dass es sich wirklich um ein absolutes Minimum handelt, da g eine quadratische Funktion zu einer nach oben geöffneten Parabel ist. Wir haben also zu einem endlichen Punkteschwarm die im Sinne des Kleinsten-Quadrate-Kriteriums optimale Gerade ausgerechnet.

Man kann das Problem auch abstrakter fassen und eine Variable $aX+b$ derart suchen, dass die Varianz der Fehlervariable E - das ist eine Funktion von a und b - minimalisiert wird, also wieder ausgehen von

$$\begin{aligned} Y &= aX + b + E, \text{ also} \\ E &= Y - aX - b, \text{ zu minimalisieren ist daher:} \\ g(a) &= \sigma^2(E) = \sigma^2(Y - aX) = \mu((Y-aX)^2) - \mu^2(Y - aX) \\ &= a^2\sigma^2(X) - 2aCov(X,Y). \end{aligned}$$

Man hat sofort

$$g'(a) = -2Cov(X,Y) + 2a\sigma^2(X).$$

Null wird das genau für

$$a = \frac{Cov(X,Y)}{\sigma^2(X)},$$

und auch das ergibt sofort - und am elegantesten - die alte Formel. Interessant ist es dabei, dass nichts über die auch nur lineare Unabhängigkeit zwischen X und E vorausgesetzt wurde, dass man diese vielmehr als Konsequenz der Minimalität von $\sigma^2(E)$ mit dem ausgerechneten Parameter a erhält.

2.2. Verallgemeinerung des Problems auf mehrere unabhängige Variablen und auf nichtlineare Regression. Wir behandeln zunächst den *linearen* Fall mit mehreren unabhängigen Veränderlichen, also die Modellgleichung:

$$Y = a_1X_1 + a_2X_2 + ... + a_nX_n + E. \tag{2.7}$$

Man beachte, dass eine additive Konstante hier überflüssig ist, weil man dafür etwa eine Variable (Erinnerung: Das klingt komisch, aber wir waren immer in der Lage, Variablen mit nur einem Wert - also Streuung Null - ebenso zu behandeln wie alle anderen!) X_n = Konstante mit Wert 1 zusätzlich zu den anvisierten Variablen hinzufügen kann. Dann ist a_n die additive Konstante. Nunmehr ist die zu minimalisierende Varianz von E eine Funktion von den n Parametern $a_1, ..., a_n$. Diese sind also nunmehr unsere unabhängigen Variablen der Funktion

$$g(a_1, ..., a_n) = \sigma^2(Y - \sum_{i=1}^{n} a_iX_i).$$

Gesucht ist die (wieder stellt sich heraus, dass es eine einzige gibt) Parameterfolge, für die das minimal wird. Es liegt nahe, dass man dazu auch die Ableitung benutzt, die wir aber für Funktionen mehrer Veränderlicher noch gar nicht kennen. Die Idee dazu können wir jedoch anhand unseres Extremwertproblems leicht einführen:

Stellen wir uns vor, für die konkrete Folge $(b_1, ..., b_n)$ hätten wir einen *lokalen* Extremwert von g, d.h. für alle $(a_1, ..., a_n)$ in einer kleinen Umgebung von $(b_1, ..., b_n)$ gälte $g(a_1, ..., a_n) \geq g(b_1, ..., b_n)$. Dann gilt *insbesondere* für a_1 in einer kleinen Umgebung von b_1, dass $g(a_1, b_2, ..., b_n) \geq g(b_1, ..., b_n)$. Das heißt aber, dass an der Stelle b_1 ein lokaler Extremwert der Funktion $h(a) = g(a, b_2, ..., b_n)$ liegt. Also können wir folgern, dass $h'(b_1) = 0$ gilt. Das ist eine Gleichung mit den Unbekannten $b_1, ..., b_n$. Dasselbe können wir nun für jede Variable von g durchführen, und so erhalten wir n Gleichungen für unsere n Unbekannten. Die Idee war also, aus einer Funktion mit n unabhängigen Veränderlichen n Funktionen mit nur einer zu machen und deren Ableitungen zu betrachten. Man variiert also immer nur eine Variable und betrachtet die anderen als konstant. Dafür hat man nun eine bequeme Notation und verbale Bezeichnung:

Partielle Ableitungen von einer Funktion $g{:}(a_1, ..., a_n) \mapsto g(a_1, ..., a_n)$:

$$\begin{aligned} \frac{\partial}{\partial a_i} g(a_1, ..., a_n) &= h_i'(a_i), \text{ wobei} \\ h_i(a_i) &= g(a_1, ..., a_n), \text{ worin nunmehr alle } a_j,\ j \neq i, \\ &\quad \text{als äußere Parameter betrachtet werden.} \end{aligned} \tag{2.8}$$

Notwendige Bedingung für das Auftreten von lokalen Extremwerten bei einer Funktion mehrerer Veränderlicher ist demnach das Verschwinden aller partiellen Ableitungen (sofern diese existieren). Übrigens errechnet man aus den partiellen Ableitungen auch die Gesamtableitung zur linearen Approximation, Existenz der letzteren ist aber noch etwas mehr als bloße Existenz der partiellen Ableitungen.

Wie sieht nun das Gleichungssystem in unserem Beispielproblem aus, das durch diesen Prozess entsteht? Glücklicherweise ist es viel einfacher als man allgemein fände, nämlich ein *lineares (*quadratisches*)* Gleichungssystem, das man stets eindeutig lösen kann, wenn nur die Gleichungen unabhängig (wieder ein anderer Begriff von Unabhängigkeit - keine Gleichung soll aus den übrigen folgen, und die Gleichungen sollen miteinander verträglich sein) sind. (Bei Daten zu Prädiktorvariablen läuft das in unserem Beispiel auf eine gewisse (das ist übrigens *wieder* eine andere Unabhängigkeit, nämlich die lineare im Sinne der Vektorraumtheorie!) Unabhängigkeit der einzelnen Variablen hinaus. Klar ist z.B., dass man nicht erwarten kann, Y eindeutig als $a_2X + a_2X$ darzustellen (hier wäre $X_1 = X_2$). Mehr und besser noch: In unserem linearen Spezialfall gelangt man mit der eindeutigen Lösung auch zu einem absoluten Minimum, also Optimum.

Das lineare Gleichungssystem für die Modellgleichung
$$\mathbf{Y = a_1X_1 + ... + a_nX_n + E}$$

Tatsächlich lässt sich diese sehr leicht ausrechnen. Man hat mit der Forderung $\mu(E) = 0$ (unter den Prädiktorvariablen sei eine Konstante untergebracht, dann ist das unproblematisch):

$$\begin{aligned} g(a_1, ..., a_n) &= \sigma^2\left(Y - \sum_{i=1}^{n} a_iX_i\right) = \mu\left(\left(Y - \sum_{i=1}^{n} a_iX_i\right)^2\right) \text{ (mit } \mu(E) = 0) \\ &= \mu\left(Y^2 - 2\sum a_iX_iY + \left(\sum a_iX_i\right)^2\right) \\ &= \mu(Y^2) - 2\sum a_i\mu(X_iY) + \sum_{i\neq j} a_ia_j\mu(X_iX_j) + \sum_{i=1}^{n} a_i^2\mu\left(X_i^2\right) \end{aligned}$$

Dies ist zu minimalisieren. Dies ergibt das System der folgenden n linearen Gleichungen, von denen die i-te lautet:

$$\frac{\partial}{\partial a_i}g(a_1, ..., a_n) = -2\mu(X_iY) + 2\sum_{i\neq j} a_j\mu(X_iX_j) + 2a_i\mu\left(X_i^2\right) = 0.$$

Ordentlich als lineares Gleichungssystem hingeschrieben, nach Division durch 2:

$$\sum_{i\neq j} a_j\mu(X_iX_j) + a_i\mu\left(X_i^2\right) = \mu(X_iY),\ 1 \leq i \leq n.$$

In der konkreteren vertrauteren Form als lineares Gleichungssystem mit Pünktchen geschrieben:

$$\begin{aligned} a_1\mu(X_1^2) + a_2\mu(X_1X_2) + a_3\mu(X_1X_3) + ... + a_n\mu(X_1X_n) &= \mu(X_1Y) \\ a_1\mu(X_2X_1) + a_2\mu(X_2^2) + a_3\mu(X_2X_3) + ... + a_n\mu(X_2X_n) &= \mu(X_1Y) \\ \vdots &= \vdots \text{ (weitere Zeilen)} \\ a_1\mu(X_nX_1) + a_2\mu(X_nX_2) + a_3\mu(X_nX_3) + ... + a_n\mu(X_n^2) &= \mu(X_nY). \end{aligned} \tag{2.9}$$

Wenn die Variablen nur endliche Population haben oder aber mit Stichproben gearbeitet wird, dann stimmt die Lösung dieses Gleichungssystems wieder mit der Kleinsten-Quadrate-Lösung überein, und das Gleichungssystem erhält die konkretere Gestalt, nach Weglassen der Faktoren $\frac{1}{n}$: Wir bezeichnen mit $x_{i,k}$ (man könnte das auch ohne Komma schreiben, das Komma dient nur zur Verdeutlichung dessen, dass es sich um einen Doppelindex handelt, nicht etwa um ein Produkt) den Wert der Variablen X_i beim k-ten Populations- bzw. Stichprobenmitglied, mit y_k den Wert der Variablen Y bei eben diesem Mitglied. Dabei laufe k von 1 bis $N =$ Populations- bzw. Stichprobenumfang. (Diese Zahl N müssen wir sorgsam unterscheiden von n, der Anzahl der benutzten unabhängigen Variablen. Normalerweise wird n sehr klein gegen N sein.)

$$\begin{aligned}
a_1\sum_{i=1}^N x_{1,i}^2 + a_2\sum_{i=1}^N x_{1,i}x_{2,i} + a_3\sum_{i=1}^N x_{1,i}x_{3,i} + ... + a_n\sum_{i=1}^N x_{1,i}x_{n,i} &= \sum_{i=1}^N x_{1,i}y_i\\
a_1\sum_{i=1}^N x_{2,i}x_{1,i} + a_2\sum_{i=1}^N x_{2,i}^2 + a_3\sum_{i=1}^N x_{2,i}x_{3,i} + ... + a_n\sum_{i=1}^N x_{2,i}x_{n,i} &= \sum_{i=1}^N x_{2,i}y_i\\
&\vdots = \vdots\\
a_1\sum_{i=1}^N x_{n,i}x_{1,i} + a_2\sum_{i=1}^N x_{n,i}x_{2,i} + a_3\sum_{i=1}^N x_{n,i}x_{3,i} + ... + a_n\sum_{i=1}^N x_{n,i}^2 &= \sum_{i=1}^N x_{n,i}y_i.
\end{aligned}$$

Manchmal wird hier zwecks größerer Übersichtlichkeit die Notation $[x_ix_j]$ für $\sum_{k=1}^N x_{i,k}x_{j,k}$ bzw. $[x_iy]$ für $\sum_{k=1}^N x_{i,k}y_k$ verwandt. In der i-ten Zeile steht also auf der rechten Seite $[x_iy]$, auf der linken bei der Unbekannten a_j der Faktor (Koeffizient) $[x_ix_j]$. Übrigens lösen Computerprogramme diese Gleichungssysteme bequem auf, nachdem man in geeigneter Weise die Vektoren $(x_{1,k}, x_{2,k}, x_{3,k}..., x_{n,k}, y_k)$, $1 \le k \le N$, eingegeben hat.

Zur *Konkretisierung* schauen wir einmal nach, was wir mit $n = 2$, $X_1 = X$, $X_2 = 1$ bekommen. Wir schreiben auch a für a_1 und b für a_2. Natürlich ist das genau das einfache Modell der linearen Regression mit einer unabhängigen Variablen, das wir oben besprachen, und selbstverständlich sollte das alte Ergebnis herauskommen. Aber an diesem kleinen Beispiel kann man schon das Funktionieren des verallgemeinerten Ansatzes beobachten. Das Gleichungssystem lautet konkret (mit den Summen, nicht den Erwartungswerten) - man beachte, dass stets $x_{2,k} = 1$, da X_2 konstante Größe mit Wert 1 ist, außerdem $x_{1,k}$ nunmehr vereinfacht x_k heißt:

$$\begin{aligned}
a\sum_{k=1}^N x_k^2 + b\sum_{k=1}^N x_k &= \sum_{k=1}^N x_ky_k\\
a\sum_{k=1}^N x_k + b\sum_{k=1}^N 1 &= \sum_{k=1}^N y_k \text{ (beachte: } \sum_{k=1}^N 1 = N).
\end{aligned}$$

Dies ist ein übersichtliches (lineares ohnehin) (2×2)-Gleichungssystem, und wir lösen es auf: Die zweite Zeile ergibt sofort

$$b = \frac{1}{N}\sum_{k=1}^N y_k - a \cdot \frac{1}{N}\sum_{k=1}^N x_k = \overline{y} - a\overline{x}.$$

Eingesetzt in die erste Gleichung:

$$a\left(\left(\sum_{k=1}^{N} x_k^2\right) - \overline{x}\sum_{k=1}^{N} x_k\right) + \overline{y}\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} x_k y_k$$

ergibt das

$$a = \frac{\sum x_k y_k - N\overline{xy}}{\sum x_k^2 - N\overline{x}^2} = \frac{\sum (x_k - \overline{x})(y_k - \overline{y})}{\sum (x_k - \overline{x})^2}.$$

Tatsächlich das alte Resultat. (Darum ging es natürlich nicht, das wurde oben eleganter hergeleitet, sondern es war nur um ein konkretes lineares Gleichungssystem zu tun, wie es sich bei mehreren unabhängigen Variablen ergibt.)

Zuvor fragten wir im Falle von nur einer unabhängigen Variablen X danach, welcher Anteil der Varianz von einer Zielvariablen Y durch lineare Regression auf X erklärt ist, und erhielten als Antwort: $\rho^2(X,Y)$ beschreibt genau diesen Anteil. Analoges ist auch bei mehreren unabhängigen Variablen vorzunehmen:

DEFINITION 18 (multipler Korrelationskoeffizient). *Der multiple Korrelationskoeffizient für die Regression von Y auf $X_1, ..., X_n$ wird mit $R(Y|X_1, ..., X_n)$ bezeichnet und ist folgendermaßen definiert: Sei*

$$Y = \sum_{i=1}^{n} a_i X_i + c + E$$

die Regressionsgleichung mit den zuvor berechneten Regressionskoeffizienten, dann ist

$$\widehat{Y} := \sum_{i=1}^{n} a_i X_i + c, \text{ manchmal ausführlicher } \widehat{Y}(X_1, ..., X_n) \text{ geschrieben,}$$

die lineare Schätzgröße für Y aus $X_1, ..., X_n$, und man definiert

$$R(Y|X_1, ..., X_n) := \rho(Y, \widehat{Y}).$$

(Das ist der gewöhnliche einfache Korrelationskoeffizient zwischen diesen beiden Variablen.)

Bemerkung zur Bezeichnung: Eigentlich sollte man großes griechisches „R" verwenden, das sieht aber so aus: P, was für Wahrscheinlichkeitsfunktion vergeben ist. Für die empirische Schätzung sollte man daher konsequent „$\widehat{R}$" verwenden, was sich wie oben unter Einsetzen der empirischen Regressionskoeffizienten $\widehat{a_i}$ mittels $r(Y, \widehat{Y}_{empir})$ (empirischer Korrelationskoeffizient, s.o. 2.1.2) ergibt, mit $\widehat{Y}_{empir} = \sum_{i=1}^{n} \widehat{a_i} X_i + c$.

SATZ 13. *Man hat mit den Bezeichnungen der Definition:*

$$R^2(Y|X_1, ..., X_n) = \frac{\sigma^2\left(\widehat{Y}\right)}{\sigma^2(Y)}.$$

Das ist der Anteil der Varianz von Y, der durch lineare Regression auf $X_1, ..., X_n$ erklärt wird.

Der Grund ist einfach der, dass die Fehlervariable E wieder linear unabhängig ist von $\widehat{Y}$ und sich somit wie im einfachen Fall die Varianz von Y zerlegt als $\sigma^2(Y) = \sigma^2(\widehat{Y}) + \sigma^2(E)$.

Verallgemeinerung auf nichtlineare Regression

Tatsächlich kann man mit dem gelösten Problem für den linearen Spezialfall auch nichtlineare Abhängigkeiten funktional darstellen. Dazu bedient man sich des einfachen Kunstgriffs, nichtlineare Funktionen $Z_i = f_i(X_i)$ der Prädiktorvariablen als neue Prädiktorvariablen Z_i zu nehmen und dann *für diese* den *linearen Ansatz*

$$Y = a_1 Z_1 + ... + a_n Z_n + E$$

zu machen. Man rechnet wie oben für den verallgemeinerten linearen Ansatz gezeigt und hat damit Y mit dem Fehler E als nichtlineare Funktion der X_i dargestellt. Selbstverständlich kann man auch allgemeiner Prädiktoren der Form $f_i(X_1, ..., X_n)$, mit einer völlig freien Anzahl dieser Funktionen f_i (also nicht notwendig n) versuchen. Natürlich besteht das Hauptproblem darin, geeignete nichtlineare Funktionen zu finden, was man keineswegs schematisieren kann. Manchmal hat man Erfolg damit, ein Polynom gewissen Grades in den $X_1, ..., X_n$ zu verwenden. Der Ansatz sieht dann z.B. bei zwei Variablen für ein Polynom zweiten Grades so aus:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_1^2 + a_4 X_2^2 + a_5 X_1 X_2 + E.$$

(Also Summe der Exponenten zu den unabhängigen Variablen ≤ 2.) In der gewöhnlichen Varianzanalyse handelt es sich übrigens genau um diesen Fall, noch abgespeckt um die Terme mit den quadrierten Variablen. Außer dem linearen Anteil steht dann lediglich noch das Produkt $X_1 X_2$ (mit einem Faktor) als „Interaktionsterm".

Bemerkungen zu Wahl und eventuell Revision des Ansatzes

Der Möglichkeiten sind viele, wörtlich unendlich viele. Aber ein generelles Prinzip ist wichtig zu beachten: Man hat Daten (Vektoren der Werte der Prädiktoren und der Zielvariablen), dazu weitere mögliche Daten aus der Population. Idee des Ganzen muss es sein, diese Datenmenge übersichtlich, möglichst einfach darzustellen. Dabei interessieren genau genommen nur die Werte der Zielgröße - die Prädiktoren und zugehörige Parameter sollen ja gerade diese Werte rekonstruieren. Das heißt aber, man sollte möglichst wenige Prädiktoren, damit möglichst wenige zu bestimmende Parameter benutzen. Dazu eine wichtige generelle Einsicht: Wenn ein einfaches Modell (also mit wenigen Parametern) recht gut auf eine Stichprobe passt, so hat man gute Aussicht, dass es auch einigermaßen zu weiteren Daten passt, die noch nicht in der Stichprobe enthalten waren. Man kann also eine recht gute Generalisierungsfähigkeit (von den Beispielen der Stichprobe auf die Gesamtpopulation) erwarten. Dagegen ist es eine Binsenweisheit, dass man mit allzu vielen Parametern eine Fülle von ganz verschiedenen Möglichkeiten hat, die Daten einer Stichprobe recht genau, ja im Prinzip beliebig genau zu reproduzieren. Aber einmal hat man damit keine gute Reduktion - das Modell selbst enthält mit den vielen Parametern viele Daten, außerdem kann man sicher sein, dass Verallgemeinerung auf weitere Daten nicht gelingt - bei denen gehen die Voraussageresultate der verschiedenen zur Stichprobe passenden Modelle dann weit auseinander. An solchen Modellen besteht demnach aus doppeltem Grund keinerlei vernünftiges Interesse. Noch eine Bemerkung: Hat man ein Modell gerechnet, so schaue man nach, welche der Koeffizienten sehr nahe bei Null liegen und damit zur Voraussage der Zielgröße kaum etwas beitragen - die kann man dann zwecks weiterer Reduktion gut weglassen. Schließlich muss noch bemerkt werden, dass die Sicherheit der Parameterbestimmung durch Stichproben (Parameterschätzung) mit demselben Problem

behaftet ist wie das Schätzen von Korrelationskoeffizienten, das wir oben illustrierten. Bei relativ kleinen Stichproben werden die Vertrauensintervalle unbrauchbar groß. Noch ein abschließender Rat: Wer zu diesem und ähnlichen Themen ernsthaft mehr als hier dargestellt lernen möchte, wende sich an amerikanische Literatur, und zwar den Teil, der selbstverständlich mit Vektoren und Matrizen arbeitet, und lerne die mathematischen Prärequisiten!

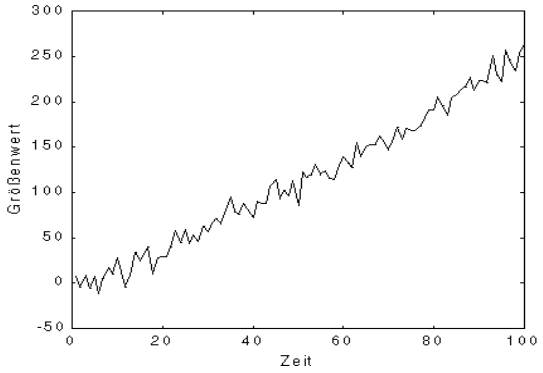
KAPITEL 6

Elementares über Zeitreihen

Verfolgt man die Entwicklung einer beobachtbaren Größe (Variablen) X in der Zeit, hält also deren Werte zu verschiedenen Zeitpunkten $t_1, ..., t_n$ fest, so beobachtet man eine Zeitreihe

$$X(t_1), ..., X(t_n).$$

(Dabei kommt es nicht darauf an, ob die Abstände zwischen den Zeitpunkten Jahre oder Millisekunden sind, nicht einmal gleich müssen sie sein.) Man beachte, dass wir damit gegenüber den vorangehenden Abschnitten ein wenig den Gesichtspunkt verändert haben: Bisher beobachteten wir typisch die Werte einer Größe Y bei verschiedenen Mitgliedern der Population: $Y(\omega_1), ..., Y(\omega_n)$. Nunmehr stellen wir uns nur ein einziges Individuum vor, an dem wir verschiedene Werte einer Größe zu verschiedenen Zeitpunkten beobachten. Das bedeutet zunächst eine weitere Struktur, da Zeitpunkte $t_1, ..., t_n$ in natürlicher Weise geordnet sind - wir werden stets $t_1 < ... < t_n$ voraussetzen, im Gegensatz zu beliebigen Individuen $\omega_1, ..., \omega_n$ einer beliebigen Population. Daher macht es stets Sinn, eine solche Zeitreihe graphisch darzustellen in der naheliegenden Form (man kann zusätzlich die Messzeitpunkte mit Punkten hervorheben, auch die Verbindungen zwischen den Messpunkten weglassen, die ohnehin nur zur leichteren Lesbarkeit dienen sollten und keine eigenständige Bedeutung haben):

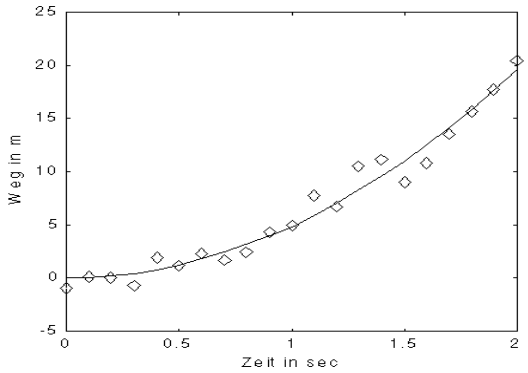


Ansonsten betrachten wir die Werte $X(t_1), ..., X(t_n)$, die zu *einem* Individuum *zu verschiedenen Zeitpunkten* gehören, durchaus völlig analog dazu, wie wir bisher Werte $Y(\omega_1), ..., Y(\omega_n)$ einer Größe *an verschiedenen Individuen* betrachteten. Wir reden also (deskriptiv) von Mittelwert und Streuung usw., doch hier gibt es wiederum Veränderungen: Wir werden mit einer Zeitreihe im allgemeinen (außer eben bei völlig zufälligen zeitlichen Prozessen) keine Zufallsstichprobe haben, anhand deren man auf die gesamte Zeitreihe (als Grundgesamtheit oder Gesamtpopulation schliessen könnte. Dem entspricht, dass wir weitere deskriptive Mittel benötigen,

um die Strukturen zu beschreiben, deren Anwesenheit den Zufallscharakter durchbricht. Letztlich strebt man immer eine Analyse einer Zeitreihe in feste Strukturen plus Zufallselement an. Ein besonders schwieriges Problem ist es im allgemeinen, eine Zeitreihe zu extrapolieren, d.h. Werte vorauszusagen, die zu späteren Zeitpunkten eintreten werden. Tatsächlich ist das Thema „Zeitreihen" ein sehr weites Feld, auf dem mittlerweile einige tiefergehende mathematische (bei weitem nicht nur wahrscheinlichkeitstheoretische!) Methoden verwandt werden, mit einigen Erfolgen - man erwarte jedoch nicht genaue Voraussagen zu den Aktienkursen der nächsten Jahre oder auch zum Wetter. - Dafür gibt es mathematische Begründungen dafür, warum bei *gewissen* Zeitreihen eine Extrapolation auf lange Zeit auch nur mit bescheidenster Genauigkeit *prinzipiell unmöglich ist.* (Stichwort Chaostheorie - das Wetter ist dabei ein Kandidat, der nach plausiblen Überlegungen dazugehört.) Wir wollen hier nur einen bescheidenen praktischen Zweck verfolgen: Gewisse Elemente, in die man Zeitreihen routinemäßig mit Gewinn zerlegen kann, wollen wir anschaulich vorstellen und deskriptiv mathematisch beschreiben und nutzen. Wir beschränken uns also auf das Verstehen wesentlicher Elemente von gegebenen Zeitreihen, gerade so, wie man in deskriptiver Statistik eine Stichprobe (oder bekannte Gesamtheit) nur kompakter zu beschreiben sucht. Wir beschränken uns hier zunächst auf die drei Elemente „Trend", „saisonaler oder (allgemeiner gesprochen) periodischer Anteil", „Zufallsanteil" (gemeint im Sinne von „unanalysierter Rest", mit der Möglichkeit, den zufälligen Charakter zu testen).

Trend

Man sagt, die Arbeitslosenzahlen hätten in den letzten Jahren einen starken Aufwärtstrend gehabt, und meint damit gerade, dass ein systematischer Anstieg (sogar ein starker) vorlag. Das bedeutet nicht, dass nicht auch einmal ein momentanes Absinken eintrat: Saisonal bedingt gab es das. Aber es bedeutet, dass bei gröberer Sicht der Anstieg das wesentliche Merkmal war, im Beispiel hätte man nur saisonal vergleichbare Zeitpunkte nehmen müssen, um über einige Jahre nichts als Anstieg wahrzunehmen. Ein anderes Beispiel: Stellen wir uns vor, dass man zu verschiedenen Zeitpunkten misst, wie weit ein losgelassener Körper gefallen ist, so bemerkt man wieder einen systematischen Anstieg. Tatsächlich wäre auch in diesem Falle das Gesetz des Anstiegs mit der Zeit stellenweise durchbrochen, wenn man sich vorstellt, dass die Messungen einen Grad von Zufalls-Ungenauigkeit hätten und die Messzeitpunkte sehr dicht beieinander lägen. Das könnte dann so aussehen:

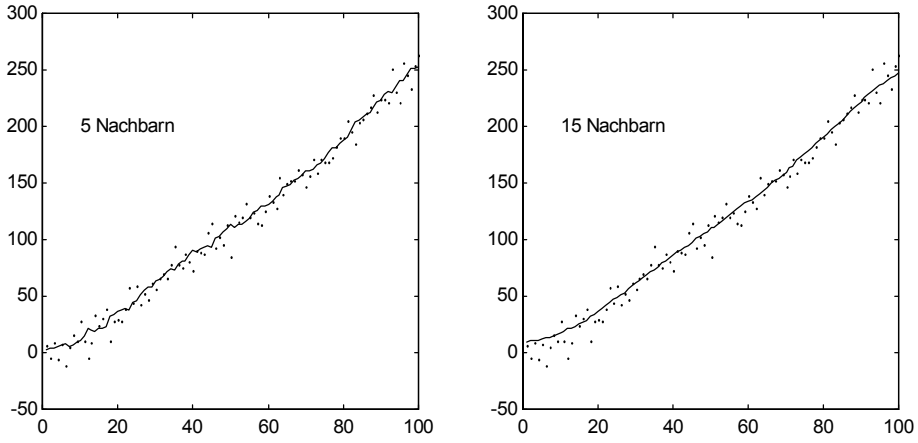


In diesem Beispiel haben wir nur den (hier exakten!) Trend gemäß physikalischer Gesetzmäßigkeit plus die Zufalls-Messfehler, keinerlei saisonale Schwankungen.

Wie kann man aus einer gegebenen Zeitreihe quantitativ den Anteil „Trend" herauspräparieren? Anschaulich legt es sich nahe, eine Zeitreihe mit Fluktuationen (ob saisonal oder/und bizarr zufällig) einfach grob zu glätten. Im Prinzip kann man eine solche glatte Kurve nach Gefühl einzeichnen. Aber es ist zumal bei Computereinsatz bequem, Glättungen besser verfügbarer Form zu erlangen. Zwei Methoden dazu werden gern benutzt:

Herausschälen eines Trends: Methode der „gleitenden Mittelwerte"

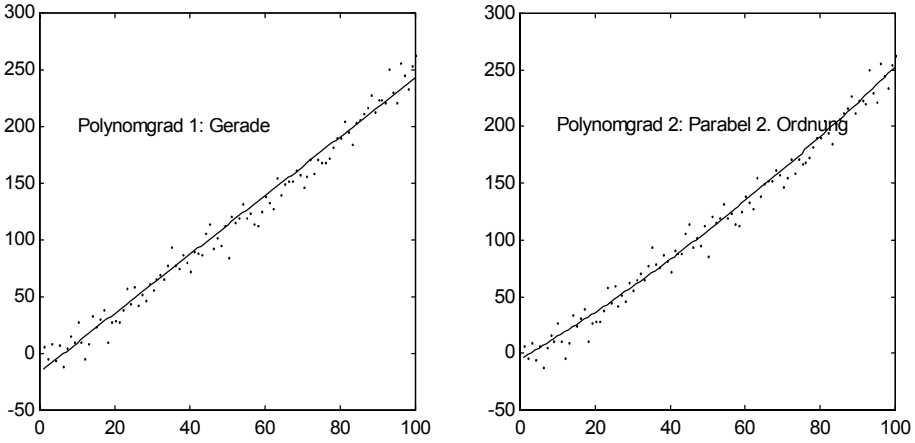
Man ersetzt dazu jeden Wert einer Zeitreihe durch den Mittelwert aller Nachbarwerte einer gewissen Umgebung (drei Nachbarn, ..., zehn Nachbarn). Natürlich muss man dabei ausprobieren, wie groß die Umgebung gewählt werden sollte. (Zu klein: Es bleibt unruhig. Zu groß: Es werden wesentliche Bewegungen unterdrückt und durch eine Konstanz ersetzt.) Ein kleines Problem hat man für die Ränder, an denen es zu einer Seite so viele Nachbarn nicht gibt: Das bewältigt man einfach, indem man die Endwerte entsprechend oft hinzufügt. Hier ist ein Beispiel, das zeigt, wie diese Methode wirkt (5 Nachbarn waren hier zu wenig, 15 sind genug):



Herausschälen eines Trends: Methode der Anpassung einer glatten Funktion

Man erreicht denselben Zweck, indem man ähnlich vorgeht, wie wenn man eine Gerade optimal in einen Punkteschwarm legt: Man nimmt eine Funktion, typisch ein Polynom wie $f(t) = a_0 + a_1 t + a_2 t^2 + ... + a_k t^k$ und paßt die Koeffizienten a_i so an, dass die Summe der Abstandsquadrate zwischen $f(t)$ und beobachtetem $X(t)$ für die Beobachtungszeitpunkte minimal wird. Dabei hat man ebenfalls darauf zu achten, dass man den Polynomgrad k nicht zu klein und nicht zu groß wählt. Zu klein: Der vorhandene Trend kann nicht nachgezeichnet werden, z.B. hat man mit $k = 0$ eine Konstante, mit $k = 0$ ein Gerade (damit könnte man nur einen linearen Aufwärts- oder Abwärtstrend beschreiben). Zu großes k würde dagegen bis dahin führen, jedes Zackeln genau nachzuzeichnen und damit keinerlei einfacheres Element herauszuschälen. (Es entsteht also ein völlig analoges Problem wie bei der vorigen Methode!)

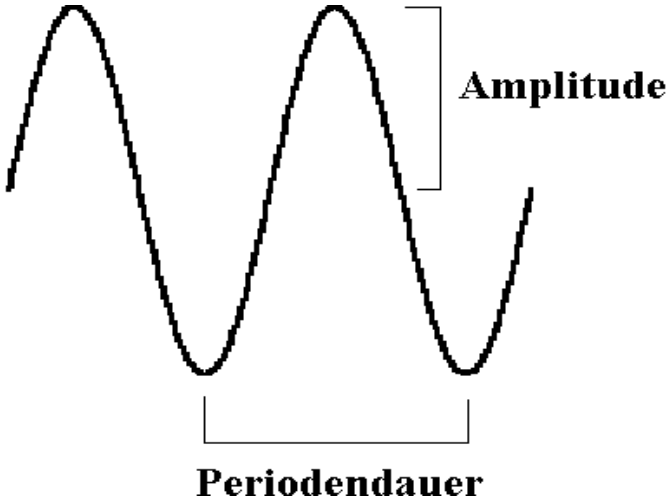
Wir illustrieren die Methode am selben Beispiel: Man sieht, dass bereits eine Gerade recht gut ist, die Parabel 2. Ordnung jedoch besser, da sie die leichte Krümmung mit erfasst:



Es sei allerdings bemerkt, dass man nicht in jedem Falle ein Polynom anpassen sollte (es gibt z.B. auch so etwas wie exponentielles Wachstum oder Annähern an einen bestimmten Wert.) Ist dergleichen systematisch zu erwarten, sollte man das nicht verkleistern mit einer Funktion falscher Form, selbst wenn diese numerisch ordentliche Ergebnisse liefert.

Saisonale Schwankungen: Idealtyp einer sinusförmigen Schwingung

Periodische Bewegungen bei Zeitreihen sind z.B. bestens vertraut als saisonale Schwankungen, wie sie bei Arbeitslosenzahlen vorkommen: Im Winter geht es herauf, im Sommer herunter. Hier ist zur Illustration eine reine saisonale Schwingung, mit Erläuterung der zugehörigen Beschreibungselemente:



Die Bewegung wiederholt sich hier in immer gleicher Form nach Ablauf einer Periodendauer. Der Ausschlag vom Mittelwert bis zum maximalen (oder minimalen) Wert, als absoluter Betrag genommen, heißt Amplitude. Gezeichnet sind im Beispiel zwei Perioden. Natürlich kann die Sache komplizierter aussehen: Einmal kann eine periodische Bewegung in sich nicht sinusförmig aussehen (das bekommt man mit Überlagerungen von Sinus- und Cosinus- Funktionen hin). Außerdem können die Amplitude oder auch die Frequenz (Anzahl der Schwingungen pro Zeiteinheit) zeitlichen Entwicklungen unterliegen - auch so etwas kann man mit etwas komplizierteren Mitteln beschreiben.

Herausschälen einer saisonalen Schwankung

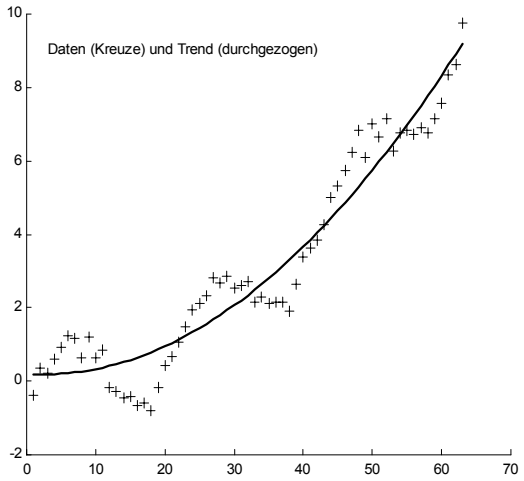
Das Einfachste: Man zieht von der vorgelegten Zeitreihe $X(t)$ den zugehörigen Trend ab. Sieht man dann eine klare saisonale Schwankung, so beschreibt man diese für sich, entweder durch Anpassen einer geeigneten Funktion, eventuell auch durch Glätten wie oben zum Trend beschrieben. Dann zieht man wieder die saisonale Schwankung ab und erhält einen unanalysierten „Rest", den man auf Zufälligkeit prüfen kann. Ein häufig auftretendes Phänomen der Nichtzufälligkeit des Restes ist dies: Es kann zwischen $X(t_i)$ und $X(t_{i+1})$ eine Korrelation bestehen, oder dasselbe mit mehreren zeitlichen Vorgängern. Eine solche Korrelation kann man durch einfache Rechnung registrieren. Dann kann man wiederum abziehen, was durch solche Korrelation erklärt ist, und es verbleibt ein geringerer unanalysierter Rest.

Analyse eine Zeitreihe in Trend, saisonale Schwankung und Zufallsrest an einem Beispiel

Wir fassen das Beschriebene kurz zu folgendem Modell zusammen:

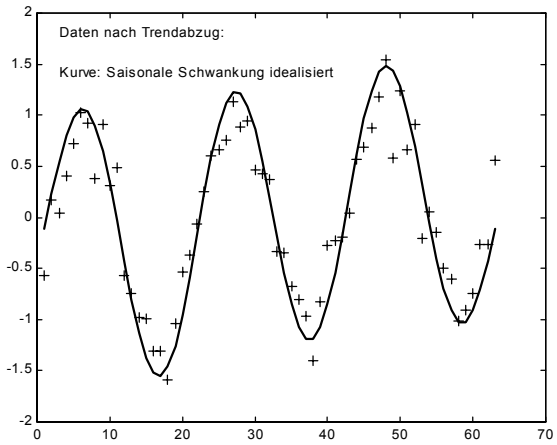
$$\begin{aligned} X(t) &= T(t) + S(t) + E(t), \text{ mit} \\ T &: \text{ Trend,} \\ S &: \text{ saisonaler Effekt,} \\ E &: \text{ unanalysierter Rest.} \end{aligned}$$

Das illustrieren wir mit folgender Zeitreihe:

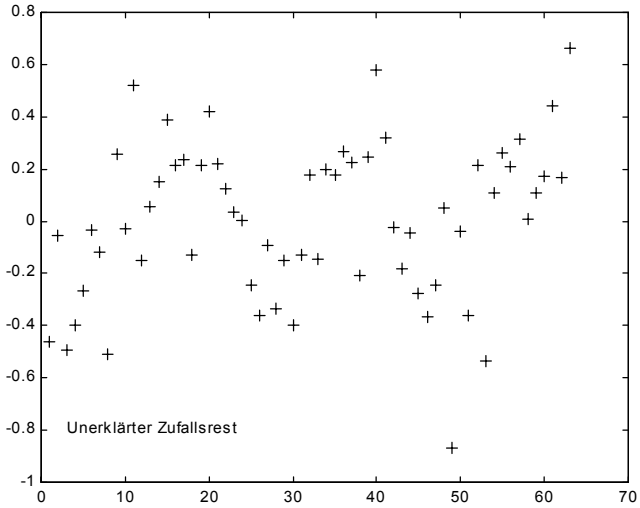


(Kreuze: Datenpunkte der Größe X, glatte Kurve: Größe T.) Man sieht einen Trend, eine saisonale Schwankung und ein „Zufallsgezackle". Ein wichtiger praktischer Anwendungszweck der Trendfunktion ist in solchen Beispielen das Herausarbeiten von saisonalen Schwankungen: Diese sind hier durchaus von respektabler Bedeutung, so dass über kürzere Zeiten ein merkliches Absinken eintritt, aber das sollte hier nicht über einen deutlichen Aufwärtstrend hinwegtäuschen. Genau in diesem Sinne spricht man etwa von „saisonbereinigten Arbeitslosenzahlen" usw.

Nach Abziehen des Trends können wir den saisonalen Anteil herausschälen - die Kreuze stellen die Daten nach Abzug der Trendfunktionswerte dar, und man sieht die Schwingung, die man mathematisch idealisierend berechnen kann ähnlich wie einen Trend, nur dass man nicht Polynome, sondern eine Summe von Sinus- bzw. Cosinusfunktionen verschiedener Perioden dafür benutzt:



(Schwingung allein: Größe S, Schwingung mit Zackeln: $S+E$). Man bemerkt in unserm Beispiel, dass der „zufällige Rest" (im Beispiel ist er wirklich als ein solcher konstruiert bzw. simuliert) quantitativ recht unbedeutend ist. Der vollständigen Übersicht halber veranschaulichen wir den Rest E noch allein, bilden also $X - T - S = E$:



(Man beachte die veränderte Skala für die Werte!)