Mathematische Statistik
(Version vom 15. 7. 2010)

Ehrhard Behrends

Fachbereich Mathematik und Informatik

Freie Universität Berlin

E i n l e i t u n g

Das Wort *Statistik* hat viele verschiedene Bedeutungen, stets geht es darum, Informationen aus der Beobachtung der uns umgebenden Welt zu ziehen. Statistische Methoden sind aus vielen Wissenschaften nicht mehr wegzudenken: Medizin, Biologie, Psychologie ... Die Methoden sind sehr vielfältig, in dieser Vorlesung wird es fast ausschließlich um Verfahren gehen, die eine mathematische Grundlage haben. Man spricht auch von *mathematischer Statistik*, dadurch soll der Ansatz von denjenigen Teilen des Gebiets abgegrenzt werden, in denen Daten nur beschrieben und übersichtlich dargestellt werden oder in denen man nicht abgesicherte Faustregeln zum Entscheidungsfinden verwendet.

Vereinfacht kann man den Unterschied zwischen Wahrscheinlichkeitsrechnung und Statistik so beschreiben: In der Wahrscheinlichkeitsrechnung betrachtet man vorgelegte Wahrscheinlichkeitsräume, und es wird dann versucht, Folgerungen für mögliche Abfragen zu ziehen. In der Statistik dagegen geht es zunächst darum, den fraglichen Raum durch Beobachtung möglichst gut zu identifizieren und zu überlegen, welche Entscheidungen man aufgrund dieser Kenntnisse treffen sollte.

Das Gebiet ist riesengroß, es besteht nicht die geringste Hoffnung, in einer einzigen Vorlesung einen systematischen Überblick zu geben. Deswegen wurde eine Auswahl getroffen, die einerseits den Interessen des Dozenten folgt und andererseits einige typische Beispiele statistischer Schlussweisen enthält. Hier der Aufbau:

- In Kapitel 0 geht es um *Vorkenntnisse*: Die sollten in dem Umfang vorhanden sein, wie sie in einer Vorlesung zur elementaren Stochastik vermittelt werden. Grundkenntnisse in Linearer Algebra sind ebenfalls unerlässlich. Vieles kann man nämlich nur dann wirklich verstehen, wenn man die Verfahren geometrisch als Aussagen über euklidische Räume deutet.

- Kapitel 1 beschäftigt sich mit *beschreibender Statistik*. Das soll recht knapp ausfallen, es geht eigentlich nur darum, einige grundlegende Begriffe und Techniken kennen zu lernen, mit denen man die bei statistischen Untersuchungen anfallenden Daten „vorbehandelt“ und veranschaulicht.

- In Kapitel 2 geht es um das *Schätzen*: Wie kann man eine Größe – z.B. einen Erwartungswert – aus den durch Beobachtung gewonnenen Daten möglichst gut schätzen? Was soll „gut“ hier überhaupt bedeuten, gibt es es eine optimale Lösung?

- Dann behandeln wir in Kapitel 3 das *Problem des Entscheidens*: Welche von verschiedenen vorliegenden Hypothesen sollte man annehmen? Wie kann man versuchen, die verschiedenen Risiken zu klassifizieren, wie sollte nach einer solchen Klassifikation eine optimale Entscheidung aussehen?

- Für Kapitel 4 sind *lineare Modelle* vorgesehen. Da geht es um Situationen, bei denen die beobachtete Größe als linearer Ausdruck in den unbekann-

ten Parametern entsteht, der dann noch von zufälligen Störungen überlagert wird. Es handelt sich um einen bemerkenswert vielseitig anwendbaren Ansatz, der viele praktisch wichtige Fragen berührt: Ist Medikament A wirkungsvoller als Medikament B? Verstärkt Rauchen die Fahruntüchtigkeit nach Alkoholkonsum? Wie hängt die Entwicklung der Intelligenz vom Fernsehkonsum ab? ...

In diesem Abschnitt wird es auch einen ausführlichen Exkurs zur mehrdimensionalen Normalverteilung geben: Die Normalverteilung spielt deswegen eine so fundamentale Rolle, weil sie einerseits in vielen konkreten Situationen auftritt und weil andererseits vergleichsweise einfache explizite Rechnungen zum Ziel führen.

Nach der Entwicklung einer allgemeinen Theorie beschäftigen wir uns insbesondere mit *Varianzanalyse* – erst danach kann man sich sinnvoll um den Vergleich von Medikamentenwirksamkeiten kümmern – und *Kovarianzanalyse* (wie beseitigt man störende Einflüsse bei der Varianzanalyse?).

- In den bisherigen Kapiteln ging es meist um das Schätzen von Zahlen oder um Entscheidungen. Eine wichtige Klasse von statistischen Problemen wird davon allerdings nicht erfasst, die so genannten *Probleme der nichtparametrischen Statistik.* Einige behandeln wir in Kapitel 5.

Daran, dass Statistik eine große gesellschaftliche Relevanz hat, wird man fast täglich bei der Zeitungslektüre erinnert. Hier einige Beispiele aus dem Frühjahr 2009:

- Lernen Jungen schwerer als Mädchen?
- Ist die neue rein-biologische Kopfschmerztablette genauso wirkungsvoll wie die aus der Chemiefabrik?
- Verführen brutale Computerspiele zur Gewalttätigkeit?
- Sind Kinder älterer Väter dümmer als andere?
- Lebt man länger, wenn man kein Fleisch isst?
- Sind die deutschen Jugendlichen ausländerfeindlich?

Ehrhard Behrends

Fachbereich Mathematik und Informatik der FU Berlin

Sommersemester 2009

Literatur: Ob man ein Lehrbuch zu einem mathematischen Thema gut oder schlecht findet, ist natürlich Geschmackssache. Ich – der Dozent – habe leider kaum eins gefunden, das ich wirklich empfehlen kann.

Am besten gefällt mir die „Stochastik“ von Georgii, in der fast alle in dieser Vorlesung behandelten Themen vorkommen. Als Ergänzung für die ersten Kapitel soll noch auf folgende Bücher hingewiesen werden: Ferguson (Mathematical Statistics), Fischer (Stochastik einmal anders), Irle (Wahrscheinlichkeitstheorie und Statistik), Krengel (Einführung in die Wahrscheinlichkeitstheorie und Statistik), Schmitz (Stochastik für Lehramtsstudenten).

Die Lehrbuchsituation zum Thema „Lineare Modelle“ ist nach meiner Einschätzung besonders unbefriedigend (etwas steht im schon erwähnten Buch von Georgii). Für die Vorbereitung habe ich noch herangezogen: Catlin (Estimation, Control, and the Discrete Kalman Filter), Christensen (Linear Models for Multivariate, Time Series and Spatial Data), Kshirsagar (A Course in Linear Models), Stapleton (Linear Statistical Models), Schach-Schäfer (Regressions- und Varianzanalyse), Toutenberg (Lineare Modelle).

Die Bücher zur Statistik findet man in der Abteilung H3 unserer Bibliothek.

Inhaltsverzeichnis

Kapitel 0

Vorkenntnisse

Es wird in dieser Vorlesung vorausgesetzt, dass die folgenden Sachverhalte bekannt sind:

Wahrscheinlichkeitsräume

- Eine *σ-Algebra* $\mathcal{E}$ auf einer Menge Ω ist eine Teilmenge der Potenzmenge, die unter allen Mengenoperationen stabil ist, bei denen höchstens abzählbar viele Elemente von $\mathcal{E}$ beteilgt sind.
- Sei $\mathcal{E}$ eine σ-Algebra auf Ω. Eine Abbildung $\mathbb{P} : \mathcal{E} \to [0,1]$ heißt ein *Wahrscheinlichkeitsmaß*, wenn $\mathbb{P}(E) = 1$ ist und

  $$\mathbb{P}(\bigcup_n E_n) = \sum_n \mathbb{P}(E_n)$$

  für jede Folge (E_n) von paarweise disjunkten Mengen in $\mathcal{E}$ gilt.
- Ein *Wahrscheinlichkeitsraum* ist ein Tripel $(\Omega, \mathcal{E}, \mathbb{P})$; dabei ist Ω eine Menge, $\mathcal{E}$ eine σ-Algebra auf Ω und $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{E})$.
- Die σ-Algebra der *Borelmengen* auf dem $\mathbb{R}^n$ ist die kleinste σ-Algebra, die alle offenen Teilmengen enthält. Faustregel: *Jede* Teilmenge, die in den Anwendungen jemals vorkommen kann, ist eine Borelmenge.

Wichtige Beispiele für Wahrscheinlichkeitsräume

- Ist Ω endlich oder höchstens abzählbar, so ist $\mathcal{E}$ in der Regel die Potenzmenge. Ein Wahrscheinlichkeitsmaß ist dann durch die Angabe der Zahlen $\mathbb{P}(\{\omega\})$ definiert. (Diese Zahlen müssen nichtnegativ sein und sich zu Eins summieren.)
- Die wichtigsten Beispiele dazu sind
  - Laplaceräume: Da ist Ω endlich, und alle Elementarereignisse haben die gleiche Wahrscheinlichkeit.

  - Bernoulliräume. Hier ist $\Omega = \{0, 1\}$, und es reicht die Angabe der Zahl $p = \mathbb{P}(\{1\})$ („Wahrscheinlichkeit für Erfolg“), um das Wahrscheinlichkeitsmaß festzulegen.
  - Abgeleitet von Bernoulliräumen sind die geometrische Verteilung (warten auf den ersten Erfolg), die Binomialverteilung (k Erfolge in n Versuchen), die hypergeometrische Verteilung (Ziehen ohne Zurücklegen) und die Poissonverteilung (Grenzwert von Binomialverteilungen).

- Sei zunächst Ω eine „einfache“ Teilmenge von $\mathbb{R}$ (etwa ein Intervall) und $f : \Omega \to \mathbb{R}$ eine „gutartige“ (etwa eine stetige) nichtnegative Funktion mit Integral Eins. Dann kann damit ein Wahrscheinlichkeitsraum durch die Festsetzung
$$\mathbb{P}(E) := \int_E f(x)\,dx$$
definiert werden. Dabei kann E eine beliebige Borelmenge sein, für die Anwendungen reicht es aber so gut wie immer, sich für E ein Teilintervall von Ω vorzustellen. f heißt dann die *Dichtefunktion* zu dem so definierten Wahrscheinlichkeitsmaß.

- Die wichtigsten Beispiele sind
  - Die Gleichverteilung auf $[a, b]$; da ist $f(x) := 1/(b - a)$.
  - Die Exponentialverteilung zum Parameter $\lambda > 0$; sie ist durch die Dichtefunktion
$$f(x) := \lambda \cdot e^{-\lambda x}$$
definiert. Durch die Exponentialverteilung kann gedächtnisloses Warten beschrieben werden.
  - Die Normalverteilungen $N(\mu, \sigma^2)$. Sie haben – für $\mu \in \mathbb{R}$ und $\sigma > 0$ – die Dichtefunktion
$$f(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$
Sie spielen für die Statistik eine ganz besonders wichtige Rolle.
  - Im Lauf der Vorlesung werden viele weitere Dichtefunktionen auftreten, die für die Statistik maßgeschneidert sind: t-Verteilungen, F-Verteilungen, χ^2-Verteilungen, ...

- Die gleiche Idee kann in allen Situatitonen ausgenutzt werden, in denen ein Integral zur Verfügung steht. Wer also auf $\mathbb{R}$ das Lebesgue-Integral kennen gelernt hat, kann integrierbare Dichten zulassen, wer die Integration im $\mathbb{R}^n$ beherrscht, kann leicht ein Wahrscheinlichkeitsmaß auf den Borelmengen dieses Raumes angeben usw. Für uns wird das später auch sehr wichtig werden, Eigenschaften mehrdimensionaler Normalverteilungen werden eine wichtige Rolle spielen.

Wahrscheinlichkeitstheorie: Grundbegriffe

- Bedingte Wahrscheinlichkeit.
- Was bedeutet „Unabhängigkeit“ für zwei, endlich viele bzw. beliebig viele Ereignisse?
- Zufallsvariable.
- Erwartungswert und Streuung.
- Unabhängigkeit für Zufallsvariable.

Grenzwertsätze

Die Grenzwertsätze besagen, „dass der Zufallseinfluss verschwindet“, wenn sich „viele“ Zufallseinflüsse unabhängig überlagern. Genauer:

- Was bedeuten „Konvergenz in Wahrscheinlichkeit“, „Konvergenz in Verteilung“, „Fast sichere Konvergenz“?
- Das Wurzel-n-Gesetz.
- Das schwache Gesetz der großen Zahlen.
- Das starke Gesetz der großen Zahlen.
- Der zentrale Grenzwertsatz.

Lineare Algebra

- Vektoren und Matrizen.
- Selbstadjungierte Matrizen und Hauptachsentransformation.
- Positiv definite Matrizen.
- Räume mit Skalarprodukt (unitäre) Räume.

Kapitel 1

Beschreibende Statistik

In der *beschreibenden Statistik* geht es darum, aus „Zufallsexperimenten“ gewonnene Daten so aufzubereiten, dass man erste Schlüsse daraus ziehen kann. In den folgenden Abschnitten sollen die wichtigsten Verfahren und Begriffe vorgestellt werden, die man dabei verwendet.

1.1 Statistische Merkmale

In diesem Abschnitt soll nur darauf hingewiesen werden, dass Daten, die interessante Aspekte darstellen, von sehr unterschiedlicher Art sein können. Formal geht es stets um Zufallsvariable $X : \Omega \to B$, wobei $(\Omega, \mathcal{E}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und B ein Messraum ist, dabei werden unabhängige Ausgaben von X beobachtet. Unterscheide:

- B ist einfach eine Menge ohne jede weitere Struktur. Dann wird es nicht sinnvoll sein, von einem „Erwartungswert“ von X zu sprechen. Das ist zum Beispiel der Fall, wenn B die Menge der Automarken oder die Menge der Grundfarben ist.

  Durch B gemessene Aspekte heißen dann *qualitative Merkmale*.

- Manchmal trägt B eine natürliche Ordnung, z.B., wenn als Elemente von B Zeugnisnoten oder Stadien des Wohlbefindens auftreten („sehr gut, gut, so la la, schlecht“). Man spricht dann von *Rangmerkmalen.* Auch hier ist es nicht sinnvoll, Erwartungswerte zu betrachten, da eine quantitative Wertung meist nicht vorliegt (und wenn, dann ist sie wie im Beispiel der Zeugnisnoten vielleicht nur vorgetäuscht).

- Es bleiben die für uns interessantesten Fälle, die *quantitativen Merkmale*. Da ist B eine Teilmenge von $\mathbb{R}$, und es wird wirklich etwas gemessen oder gezählt. Manchmal unterscheidet man noch *diskrete* und *kontinuierliche* Merkmale, je nachdem, ob B ein Intervall in $\mathbb{N}$ (evtl. nach Skalierung) oder in $\mathbb{R}$ ist. Erst hier sollte man die Frage stellen, wie denn Erwartungswerte und Streuung aussehen.

1.2 Tabellen

Es sei $X : (\Omega, \mathcal{E}, \mathbb{P}) \to B$ eine Zufallsvariable. Gegeben sind n unabhängige Kopien, die werden abgefragt, und als Ergebnis erhält man $b_1, \ldots, b_n$. So etwas heißt dann einen *Stichprobe*. In vielen Fällen ist es dabei so, dass ein $b \in B$ ein komplexes Objekt ist, das verschiedene quantitative, Rang- und qualitative Merkmale über die beobachtete Situation enthält. Bei einer Milchprobe könnten das Lieferant, Menge, Fettgehalt und Geschmack sein.
Es ist naheliegend, diese Daten in einer Tabelle, der so genannten *Urliste*, zusammenzustellen:

Nummer	Lieferant	Menge	Fettgehalt	Geschmack
1	Fa. X	4000	0.032	gut
2	Fa. Y	2500	0.040	sehr gut
⋮	⋮	⋮	⋮	⋮

Im Allgemeinen ist so eine Tabelle wenig aussagekräftig, und deswegen fasst man gewisse Aspekte der Daten auf geeignete Weise zusammen. Geht es zum Beispiel um ein quantitatives Merkmal, das reelle Zahlen in einem Intervall $[\,a, b\,]$ annehmen kann, so kann man das Intervall in r gleiche Teile teilen und zählen, wieviele der Messwerte in die einzelnen Teilintervalle fallen. So könnte eine Auskopplung aus der obigen Tabelle zum Beispiel so aussehen:

Fettgehalt	Anzahl
0.010 bis 0.019	3
0.020 bis 0.029	12
0.030 bis 0.039	4
0.040 bis 0.049	1

Damit man daran möglicherweise etwas Interessantes sieht, muss ein Kompromiss gefunden werden: Ist die Unterteilung zu fein (r also zu groß), so stehen rechts recht kleine Zahlen, die von den zufälligen Schwankungen der Messungen stark beeinflusst sind. Ist die Einteilung zu grob, so wird die Darstellung ebenfalls wenig aussagekräftig.

Machmal gibt es interessante Zusammenhänge zwischen zwei Merkmalen M_1 und M_2, die durch eine Tabelle deutlich werden sollen. Man unterteilt dazu die möglichen Werte von M_1 bzw. M_2 in endlich viele Klassen[1)] und zählt dann, wieviele der Messungen in die einzelnen Kategorien fielen. Sind etwa die M_1-Werte disjunkt in die Intervalle $I_1, \ldots, I_s$ und die M_2-Werte in $J_1, \ldots, J_t$ zerlegt, so soll n_{ij} die Anzahl der Messungen sein, bei denen M_1 in I_i und M_2 in J_j gefunden wurde. Das ergibt eine $(s \times t)$-Matrix, man spricht von einer *Kontingenztafel*:

[1)] Wieder muss ein Kompromiss gefunden werden: nicht zu grob, nicht zu fein.

	J_1	J_2	...	J_t
I_1	n_{11}	n_{12}	...	n_{1t}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_s	n_{s1}	n_{s2}	...	n_{st}

Es ist dann klar, wie die Zeilen- bzw. Spaltensummen zu interpretieren sind und dass die Summe über alle Einträge gleich der Gesamtzahl der Messungen sein muss. Betrachten wir zum Beispiel den Zusammenhang „Geschlecht“ und „Rauchgewohnheiten“. Zugrunde gelegt wird eine Umfrage unter 1000 Passanten am Kurfürstendamm. Die Urliste könnte dann so aussehen:

Nummer	Geschlecht	Raucher/in?
1	m.	ja
2	m.	nein
3	w.	nein
4	w.	nein
5	w.	ja
6	m.	ja
7	w.	nein
$\vdots$	$\vdots$	$\vdots$

Viel aussagekräftiger ist sicher die folgende Kontingenztafel zur Veranschaulichung der beiden hier interessierenden Merkmale:

	Raucher	Nichtraucher
männlich	188	292
weiblich	310	210

Teilt man jeden Tabelleneintrag noch durch die Gesamtanzahl n, so erhält man Zahlen, die bei großem n als Approximation von gewissen Wahrscheinlichkeiten aufgefasst werden können. Später werden wir das als Ausgangspunkt nehmen, um Unabhängigkeit oder gegenseitige Beeinflussung derartiger Merkmale zu studieren.

1.3 Grafische Darstellungen

Das Auge kann Zusammenhänge oft sehr effektiv erkennen, und deswegen sind grafische Darstellungen von großer Wichtigkeit. Mathematisch Bemerkenswertes lässt sich zu den Tortendiagrammen und Histogrammen allerdings nicht sagen, außer natürlich: Die Aufteilung des Laufbereichs der Merkmale darf nicht zu fein und nicht zu grob sein, um informative Grafiken zu erhalten. An dieser Stelle könnte man schon etwas zum Thema „Lügen mit Statistik“ sagen[2)].

[2)] Ich empfehle allen die gut geschriebenen Bücher von W. Krämer zu diesem Thema.

Histogramme kann man natürlich auch mathematisch formal behandeln. Sind $x_1, \ldots, x_n$ Punkte in einem Intervall $[a, b]$, ist dieses Intervall disjunkt in $I_1, \ldots, I_s$ zerlegt und bezeichnet man für eine Teilmenge I von $\mathbb{R}$ die zugehörige *charakteristische Funktion* mit χ_I, so ist das entsprechende Histogramm bis auf Skalierung durch die Funktion

$$x \mapsto \sum_{k=1}^{n} \sum_{i=1}^{s} \chi_{I_i}(x)\chi_{I_i}(x_k)$$

gegeben. So eine Darstellung ist eigentlich nur dann sinnvoll, wenn man die Zeichnung einem Computer anvertrauen möchte oder wenn man begründen muss, dass sich bei messbaren I_i messbare Funktionen ergeben.

1.4 Stichprobenmittel und -varianz, Median

Gegeben seien n Zahlen, die den Messungen $x_1, \ldots, x_n$ eines quantitativen Merkmals in einer Stichprobe entsprechen. In diesem Abschnitt geht es um einige Maßzahlen, die sich zur groben Beschreibung bewährt haben. Sie dienen zur Einschätzung gewisser typischer Aspekte der Situation. Es handelt sich wirklich um sehr grobe Beschreibungshilfsmittel, und oft lassen sich leicht Gegenbeispiele finden, wo in konkreten Situationen gerade nicht das dargestellt wird, was man eigentlich beschreiben möchte.

Stichprobenmittel

Das *Stichprobenmittel* ist einfach der Mittelwert der Messwerte, er wird mit $\overline{x}$ (gesprochen „x quer“) bezeichnet:

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

Die Idee dabei: Das Stichprobenmittel soll eine Einschätzung davon geben, wie groß die x_i „im Mittel“ sind. Das wird in den meisten Fällen auch wirklich erreicht, manchmal ist es aber kein geeignetes Maß:

- Wenn 10 Mitarbeiter einer Firma befragt werden, davon 9 jeweils 2000 Euro verdienen und ein einziger (Aufsichtsrat!) 22000 Euro bekommt, so ist der Mittelwert 4000 Euro. Das dient sicher nicht dazu, über die Gehaltsstruktur etwas Sinnvolles auszusagen.

- Genau so könnte man in einem Dritte-Welt-Land nach einer Umfrage (viele Arme, einige sehr Reiche) zu dem beruhigenden Ergebnis kommen, dass es den Leuten doch gar nicht so schlecht geht.

- In einem physikalischen Labor soll die Erdbeschleunigung durch Experimente ermittelt werden. Wenn dann einige Ausreißer dabei sind (grobe Ablesefehler, die U-Bahn fährt vorbei, ...), so wird der Mittelwert sicher kein guter Ausgangspunkt für eine Präzisionsmessung sein.

Stichprobenvarianz

Das Stichprobenmittel ist natürlich die statistische Variante des Erwartungswertes aus der Wahrscheinlichkeitsrechnung. Nun kommen wir zum Analogon der Varianz. Wieder geht es darum, ein Maß dafür zu finden, wie stark die Werte um den Mittelwert streuen. Dafür gibt es viele Möglichkeiten, man könnte die Größe des Vektors

$$(x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_n - \overline{x})$$

je nach Problemstellung in verschiedenen Normen des $\mathbb{R}^n$ berechnen. Als wichtigste Maßzahl hat sich dabei die Wahl der euklidischen Norm herausgestellt, bei ihr werden Messwerte in der Nähe des Mittels wenig gewichtet, Abstände, die größer als Eins sind, gehen besonders stark in die Berechnung ein. Wie in der Wahrscheinlichkeitsrechnung ist die Bevorzugung der quadratischen Wichtung der Abweichung eher pragmatisch als logisch zu begründen.

In der nun folgenden exakten Definition wird bei der Mittelwertbildung durch $n-1$ geteilt, man hätte hier eigentlich eher die Zahl n erwartet[3)]: Unter der *Stichprobenvarianz* (auch: *empirische Varianz*) der Stichprobe $x_1, \ldots, x_n$ versteht man die Zahl

$$V_x := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Erwartungsgemäß ist dann die *Stichproben-Streuung* – sie wird mit s_x bezeichnet – die Wurzel aus der Stichprobenvarianz.

Median

Um die gröbsten Probleme auszugleichen, die sich bei der Betrachtung des Stichprobenmittels als Maß für das mittlere Verhalten ergeben können, wird eine weitere Maßzahl studiert, der *Median*. Man sollte besser von *einem* Median sprechen, hier die Definition: Eine Zahl x heißt ein *Median* der Stichprobe, wenn mindestens die Hälfte der x_i größer gleich x und gleichzeitig mindestens die Hälfte der x_i kleiner gleich x ist. Sind etwa alle x_i verschieden und ist n ungerade, so gibt es einen eindeutig bestimmten Median, nämliche „das mittlere“ x_i; ist n gerade, so bilden die Mediane ein Intervall.

Man kann den Median durch Approximationen charakterisieren. Bei dieser Gelegenheit liefern wir auch eine Charakterisierung des Stichprobenmittels nach:

Satz 1.4.1.

(i) x sei eine reelle Zahl. Dann ist x genau dann ein Median der Stichprobe $x_1, \ldots, x_n$, wenn $\sum_{i=1}^{n} |x_i - x|$ minimal ist.

(ii) Für $x \in \mathbb{R}$ gilt bei vorgelegter Stichprobe $x_1, \ldots, x_n$: Es ist genau dann x gleich dem Stichprobenmittel $\overline{x}$, wenn $\sum_{i=1}^{n} (x_i - x)^2$ minimal ist.

[3)] Der Grund wird in Kapitel 2 klar werden.

Beweis: (i) Wir betrachten die Funktion

$$\varphi : x \mapsto \sum_{i=1}^{n} |x_i - x|.$$

φ ist stetig und geht für $|x| \to \infty$ gegen Unendlich, folglich wird das Minimum angenommen.

Sei x eine Minimalstelle. Es gebe a bzw. b bzw. c Indizes i mit $x_i = x$ bzw. $x_i < x$ bzw. $x_i > x$. Wenn man dann (mit einem positiven kleinen ε) von x zu $x-\varepsilon$ übergeht, so verändert sich φ um den Wert $\varepsilon a - \varepsilon b + \varepsilon c$. Da x Minimalstelle war, heißt das $\varepsilon(a-b+c) \geq 0$. Entsprechend folgt beim Betrachten von $\varphi(x+\varepsilon)$, dass $\varepsilon(a+b-c) \geq 0$. Aus den beiden Ungleichungen $a+b-c, a-b+c \geq 0$ folgt dann sofort, dass mindestens die Hälfte der x_i links und ebenfalls mindestens die Hälfte rechts von x liegt.

Damit ist gezeigt: Jeder minimierende Wert ist ein Median, und da es Minimalwerte gibt und φ auf der Menge der Mediane konstant ist, ist auch die Umkehrung bewiesen.

(ii) Man kann diese Beziehung sehr elementar mit Hilfe der Differentialrechnung zeigen (Ableitung Null setzen usw.). Zur Übung der entsprechenden Methoden führen wir den Beweis aber im Rahmen der Theorie der euklidischen Räume. Wir arbeiten im $\mathbb{R}^n$ mit der euklidischen Norm, genauer: Das Skalarprodukt wird durch

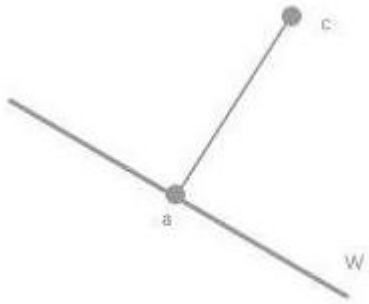
$$\langle (x_i), (y_i) \rangle := \frac{1}{n} \sum x_i y_i$$

definiert. Es sei nun $x \in \mathbb{R}$. Betrachtet man die Vektoren $a := (\overline{x}, \overline{x}, \ldots, \overline{x})$, $b := (x, x, \ldots, x)$ und $c := (x_1, \ldots, x_n)$, so steht $b - a$ senkrecht auf $a - c$; das folgt sofort aus der Definition von $\overline{x}$. Nach dem Satz von Pythagoras heißt das

$$||b - c||^2 = ||(b - a) + (a - c)||^2 = ||b - a||^2 + ||a - c||^2.$$

Wenn man das ausschreibt, steht sofort da, dass die quadratische Abweichung nur durch das Stichprobenmittel minimiert wird.

> Geometrisch kann man sich's so vorstellen: Ist $W \subset \mathbb{R}^n$ der Unterraum der konstanten Vektoren, so ist der Vektor a das Element bester Approximation in W an c. *Daher* die Orthogonalität.



□

Mediane sind wesentlich stabiler gegen „Ausreißer“ als das Stichprobenmittel. In Situationen, in denen derartige Verfälschungen zu befürchten sind, ist eine Bewertung durch den Median daher realistischer.

Als kleine Anekdote aus unserem Fachbereich ist in diesem Zusammenhang zu berichten, dass die mittlere *Studiendauer* früher durch den Mittelwert der Studienzeiten der Absolventen gemessen wurde. Da es einige Super-Langzeit-Studenten gab, führte das zu beschämend schlechten Werten. Irgendwann konnte dann die Universitätsspitze davon überzeugt werden, dass der Median ein realistischeres Maß ist. Prompt wurde die mittlere Studiendauer um zwei Semester reduziert, wir waren irgendwo ins Mittelfeld gerutscht.

1.5 Korrelation, Regressionsgerade

In diesem Abschnitt geht es um den Versuch, Zusammenhänge zwischen zwei quantitativen Merkmalen zu messen. Steigt der Ernteertrag mit der Düngemittelzugabe? Nimmt die Reisefreudigkeit mit der Arbeitslosigkeit ab? Der *Korrelationskoeffizient* ist ein sehr grobes Hilfsmittel, um dazu Informationen zu bekommen.

Die Idee ist einfach. Die Stichproben x_i und y_i für zwei Merkmale seien vorgelegt, wie üblich bezeichnen wir die Mittelwerte mit $\overline{x}$ und $\overline{y}$. Nun betrachten wir die Produkte $p_i := (x_i - \overline{x})(y_i - \overline{y})$. Dann gilt doch:

- Wenn das Verhalten von x_i mit dem von y_i nichts zu tun hat, dann wird p_i positive und negative Werte annehmen, es wird keine bevorzugte Tendenz geben.
- Ist dagegen x_i in der Regel dann groß (bzw. klein), wenn das auch für y_i gilt, so sind die p_i von der Tendenz her eher positiv.

- Haben die x_i und die y_i eine eher gegensätzliche Tendenz (x_i ist groß wenn y_i klein ist und umgekehrt), so sind die p_i meist negativ.

Zusammengefasst heißt das, dass maximal mögliche Werte von $\sum p_i$ für eine starke positive Beeinflussung sprechen, minimale Werte bedeuten einen gegensätzlichen Verlauf und Werte in der Nähe von Null können meist so gedeutet werden, dass die x'e mit den y's nichts zu tun haben.

Da man eine Größe haben möchte, die maßstabsunabhängig ist, wird noch entsprechend geteilt, die genaue Definition steht in

Definition 1.5.1. *$x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ seien quantitave Merkmale. Unter dem* Korrelationskoeffizienten *versteht man dann die Zahl*

$$r_{xy} := \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}.$$

Dabei wird angenommen, dass nicht alle x_i gleich $\overline{x}$ und nicht alle y_i gleich $\overline{y}$ sind: Dann ist der Nenner nicht Null.

Bemerkungen:

1. r_{xy} hat eine *geometrische Interpretation*. Wir betrachten wieder das am Ende des vorigen Abschnitts auf dem $\mathbb{R}^n$ eingeführte Skalarprodukt. Zunächst gehen wir von (x_i) zu $(x_i - \overline{x})$ und entsprechend von (y_i) zu $(y_i - \overline{y})$ über, nehmen also o.B.d.A. an, dass die jeweiligen Stichprobenmittel verschwinden. Dann ist r_{xy} gerade der Quotient

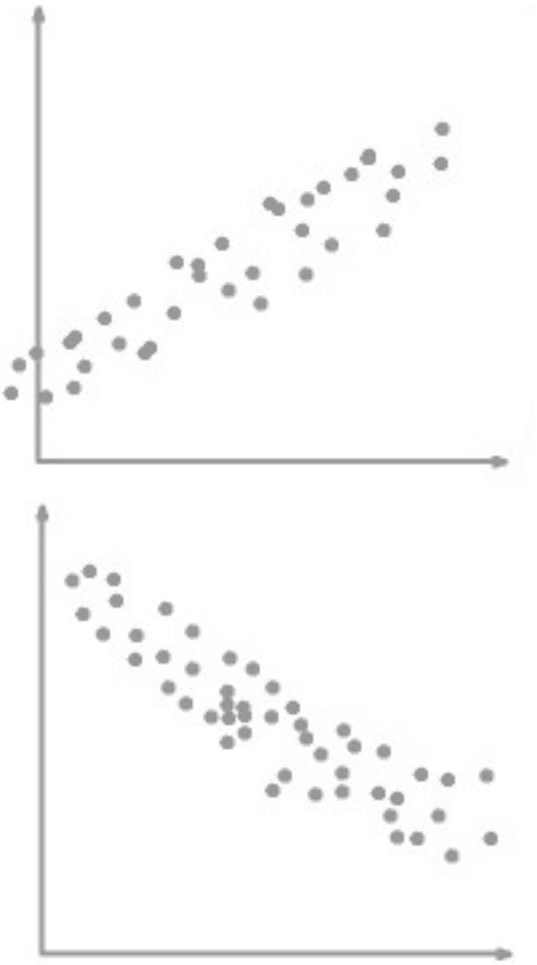
$$\frac{\langle (x_i), (y_i) \rangle}{||(x_i)|| ||(y_i)||},$$

der Faktor $1/n$ hat sich in Zähler und Nenner weggehoben. Dieser Ausdruck ist ein guter alter Bekannter. Man weiß aus der linearen Algebra:

- Er liegt zwischen -1 und $+1$, das folgt aus der Cauchy-Schwarzschen Ungleichung. Der zugehörige Arcuscosinus wird als Winkel zwischen (x_i) und (y_i) interpretiert.
- Der Wert $+1$ (bzw. -1) wird genau dann angenommen, wenn $y_i = ax_i$ für alle i und ein geeignetes $a \geq 0$ ($a \leq 0$) gilt. Das ist genau dann der Fall, wenn die Tupel (x_i, y_i) auf einer Geraden durch den Nullpunkt mit Steigung ≥ 0 bzw. ≤ 0 liegen[4)].

2. Nach der Vorrede sollte klar sein: Ist r_{xy} in der Nähe von 1, so haben die x_i die gleiche Tendenz wie die y_i zum Wachsen oder Fallen, für $r_{xy} \approx -1$ liegt eine gegensätzliche Tendenz vor, und Unabhängigkeit sollte zu $r_{xy} \approx 0$ führen. Im ersten bzw. zweiten Fall spricht man von einer *positiven bzw. negativen Korrelation.* Vorstellen kann man sich das so:

[4)] Dabei ist „durch den Nullpunkt“ zu streichen, wenn man beliebige Situationen – also nicht notwendig $\overline{x} = \overline{y} = 0$ – betrachtet: Die Approximationsmöglichkeit durch eine Gerade wird durch Verschieben des Koordinatensystems ja nicht beeinflusst.



Liegt r_{xy} in der Nähe von $+1$ oder -1, kann man versuchen, den linearen Zusammenhang zwischen den x_i und den y_i etwas genauer zu untersuchen. Das Problem stellt sich so:

> Gegeben seien (x_i, y_i) für $i = 1, \ldots, n$, man kann sich diese Menge als „Punktwolke“ in der Ebene vorstellen. Finde eine Gerade, also eine Funktion der Form $x \mapsto a + bx$, die sich dieser Punktwolke „möglichst gut anpasst“.

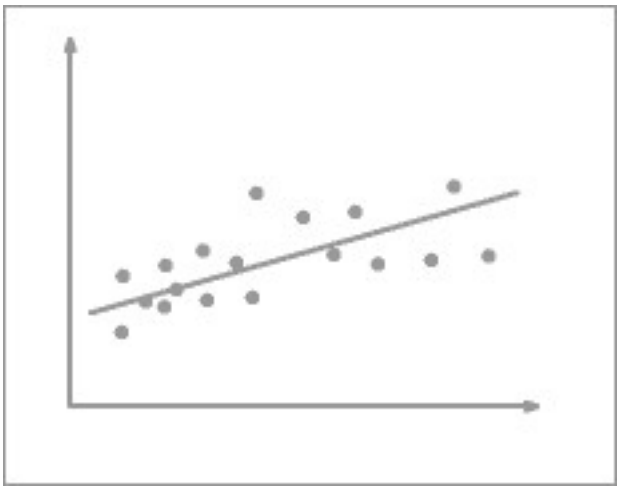
Die Anführungszeichen deuten schon an, dass noch Präzisierungsbedarf besteht, was soll *möglichst gut* heißen? Mit dem folgenden Ansatz lässt es sich am besten arbeiten:

Definition 1.5.2. *Eine Gerade $x \mapsto a + bx$ heißt eine* Regressionsgerade*, wenn*

der quadratische Abstand zur Punktwolke so klein wie möglich wird, wenn also

$$\sum_i \big(y_i - (a + bx_i)\big)^2$$

unter allen möglichen Wahlen von a, b minimal ist.



Es gibt Regressionsgeraden, und sie sind eindeutig bestimmt:

Satz 1.5.3. *Die (x_i) und die (y_i) seien gegeben, und es sei o.B.d.A. $\overline{x} = \overline{y} = 0$; das lässt sich durch eine Koordinatentransformation leicht erreichen. Dann gibt es eine eindeutig bestimmte Regressionsgerade $a + bx$, sie ist durch $a = 0$ und*

$$b = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = r_{xy} \frac{s_y}{s_x}$$

gegeben.

Beweis: Zunächst sollte man sich daran erinnern, wie man Extremwertaufgaben in mehreren Veränderlichen löst. Wir definieren

$$\varphi(a, b) := \sum_i \big(y_i - (a + bx_i)\big)^2,$$

gesucht ist ein Minimum von φ auf dem $\mathbb{R}^2$. Nun geht φ für $a, b \to \infty$ gegen Unendlich[5)]. Folglich wird das Minimum aus Stetigkeitsgründen angenommen, wir können es dadurch finden, dass wir Punkte suchen, an denen die partiellen Ableitungen von φ nach a und nach b gleichzeitig verschwinden. Man rechnet

[5)] Hier wird gebraucht, dass es mindestens zwei verschiedene x_i-Werte gibt; wir wollen das voraussetzen, sonst ist die Suche nach einer Geraden ja auch nicht sehr sinnvoll.

leicht aus: $\partial\varphi/\partial a = 0$ ist gleichwertig zu $\overline{y} = a + b\overline{x}$, und $\partial\varphi/\partial b = 0$ lässt sich zu

$$\sum_i x_i y_i = a \sum_i x_i + b \sum_i x_i^2$$

umformen. Diese beiden Gleichungen haben eine eindeutig bestimmte Lösung, nämlich die, die im Satz angegeben ist. Aus der Problemstellung ist klar, dass es sich um ein Minimum handelt[6)], die Eindeutigkeit wurde mitbewiesen. □

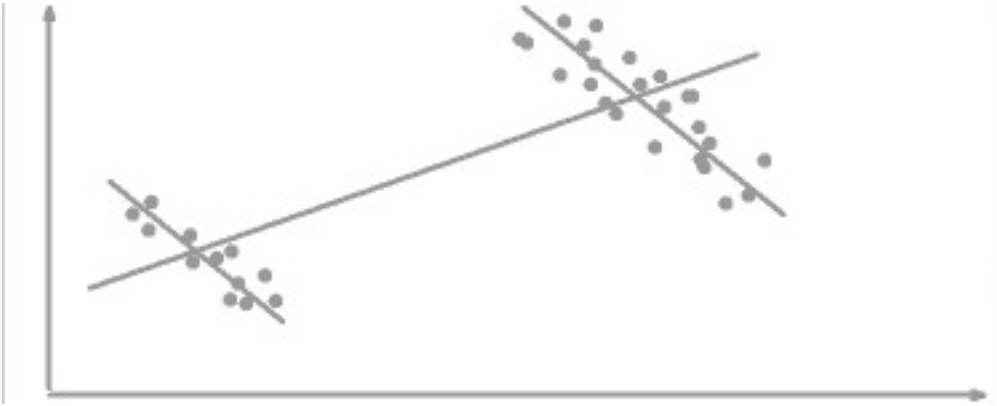
Achtung: Paradoxien

Hat eine Regressionsgerade zum Beispiel eine positive Steigung, so wird das oft so interpretiert, dass eine Zunahme des x-Merkmals „in der Regel“ eine Zunahme des y-Merkmals implizieren sollte. Hier gibt es viele Fallen, berühmt ist das *Simpson-Paradoxon*.

Zur Illustration betrachten wir die folgende Punktwolke. Sie könnte entstanden sein als Menge der Tupel

(Studiendauer, Anfangsgehalt)

bei einer Umfrage unter Universitätsabsolventen der Mathematik:



Dabei betrifft die linke „Wolke“ Bachelorabsolventen und die rechte Diplomkandidaten. Wenn man nun eine Regressionsgerade durch die Menge *aller* Tupel legt, so gibt es eine positive Steigung. Fazit:

Je länger man studiert, um so höher ist das Anfangsgehalt.

In Wirklichkeit ist es aber gerade umgekehrt. Wenn man nur die Bachelorkandidaten oder nur die Diplomkandidaten betrachtet, so sieht man, dass Langzeitstudenten eher ein geringeres Anfangsgehalt bekommen.

Ein ähnliches Paradoxon gibt es für bedingte Wahrscheinlichkeiten. Ausgangspunkt ist die folgende Tatsache aus der Bruchrechnung:

[6)] Man kann es auch streng einsehen: Die Hessematrix ist nämlich diagonal mit Einträgen $2n$ und $2\sum_i x_i^2$, ist also positiv definit.

$$\text{Aus } a_1/b_1 < x_1/y_1 \text{ und } a_2/b_2 < x_2/y_2 \text{ folgt } \textit{nicht}$$
$$(a_1 + a_2)/(b_1 + b_2) < (x_1 + x_2)/(y_1 + y_2)\,.$$

(Hier ein Gegenbeispiel: Es ist $1/4 < 101/400$ und $199/300 < 2/3$, aber $200/304$ ist größer als $103/403$.) Das kann für die Statistik wichtig sein. Bewerben sich etwa b_i Männer und y_i Frauen für das Studienfach S_i (i=1,2) und haben a_i Männer bzw. x_i Frauen Erfolg, so kann folgende Situation eintreten:

> Es ist $a_1/b_1 < x_1/y_1$ und $a_2/b_2 < x_2/y_2$, d.h. in S_1 und S_2 ist die Quote der erfolgreichen Männer schlechter als die der Frauen. Trotzdem gilt $(a_1+a_2)/(b_1+b_2) > (x_1+x_2)/(y_1+y_2)$, d.h. der Anteil der Erfolgreichen ist bei den Männern höher als bei den Frauen.

Eine Ergänzung: Ein Schritt in die Nichtlinearität

Durch die Regressionsgerade sollte doch ein *linearer* (eigentlich: affiner) Zusammenhang aufgedeckt werden: Wenn „in Wirklichkeit“ $y = a + bx$ gilt, aber nur fehlerbehaftete Messungen von y bei verschiedenen x zur Verfügung stehen, wie kann man dann sinnvolle Kandidaten für a und b finden?

Konstruktion und Beweis sind eindeutig auf den linearen Fall zugeschnitten. Trotzdem ist es nicht schwer, die gleichen Überlegungen auch auf gewisse nichtlineare Situationen zu übertragen.

Exponentielles Wachstum

Mal angenommen, es liegen Messpunkte (x_i, y_i) vor, die – wenn man sie als Punktwolke skizziert – an eine Exponentialfunktion erinnern. So etwas passiert häufig, wenn Wachstums- oder Zerfallsvorgänge beobachtet werden. Wie lassen sich dann a und b so bestimmen, dass die Kurve

$$y = a \cdot e^{bx}$$

eine möglichst gute Approximation darstellt?

Dazu wird die Gleichung $y = a \cdot e^{bx}$ zu

$$\ln y = \ln a + bx$$

umgeformt. Folglich reicht es, die Punktwolke $(x_i, \ln y_i)$ mit den bekannten Methoden durch eine Regressionsgerade $\alpha+\beta x$ zu approximieren und dann $a := e^{\alpha}$ und $b := \beta$ zu setzen[7)].

Wachstum wie bei einer Potenz x^r

Falls eine Modellierung durch $a \cdot e^{bx}$ zu ungenau ist, kann man es mit $y = a \cdot x^b$ versuchen. Auch hier findet man a und b nach einer Umformung: $y = ax^b$ ist gleichwertig zu $\ln y = \ln a + b \ln x$. Man muss also nur eine Regressionsgerade $\alpha + \beta x$ für die Paare $(\ln x_i, \ln y_i)$ finden und dann wieder $a := e^{\alpha}$ und $b := \beta$ setzen[8)].

[7)] Wenn man es „von Hand“ machen möchte, empfiehlt es sich, die (x_i, y_i) in einfach logarithmisches Papier einzutragen.

[8)] Um eine Vorstellung darüber zu bekommen, ob so ein Modell aussichtsreich ist, empfiehlt sich die Skizzierung der $(\ln x_i, \ln y_i)$ in doppelt-logarithmischem Papier. Die Punkte sollten dann ungefähr auf einer Geraden liegen.

Schlussbemerkung zur Regression: Das Thema wird in Kapitel 3 noch einmal aufgegriffen werden, nachdem die Theorie der linearen Modelle entwickelt worden ist.

Kapitel 2

Schätztheorie

In diesem Abschnitt geht es darum, mit Hilfe des Zufalls Zahlen zu schätzen. Genauer: Aus einer bekannten Familie von „Zufallsautomaten“ wird einer ausgewählt und – evtl. mehrfach – abgefragt. Aus den dann vorliegenden Informationen soll man dann eine mit der Parametrisierung der Zufallsautomaten zusammenhängende Größe schätzen, und zwar „möglichst gut“.

In diesem Zusammenhang sind einige Begriffe zu präzisieren, das geschieht in Abschnitt 2.1: Was ist ein *statistisches Modell*, was ist ein *Schätzer*? In Abschnitt 2.2 wird dann gesagt, wie man unter den vielen möglichen Schätzfunktionen eine Klassifizierung vornehmen kann: Was ist ein „optimaler Schätzer“? Bemerkenswerterweise kann man in vielen wichtigen konkreten Fällen *optimale Schätzer wirklich explizit bestimmen*, das ist der Inhalt von Abschnitt 2.3.

Die in den bisherigen Abschnitten behandelten Schätzmethoden stellen nur einen Teil der in der Statistik wichtigen Verfahren dar. In Abschnitt 2.4 gibt es noch einige *Ergänzungen zum Thema „Schätzen“*, da sollen weitere Methoden vorgestellt werden.

Für den *Spezialfall normalverteilter Messdaten* lassen sich sehr präzise Aussagen machen, dieser Fall wird in Abschnitt 2.5 näher untersucht werden.

2.1 Schätzen: Die Problemstellung

Man kann es sich so vorstellen: Hinter einem Vorhang stehen Zufallsautomaten, die durch gewisse Parameter charakterisiert sind. Einer wird ausgewählt und abgefragt, und wir sollen mit den so gewonnen Daten etwas anfangen. Etwas seriöser würde man von einer Familie von Wahrscheinlichkeitsräumen sprechen, es ist aber praktisch[1)], davon auszugehen, dass alle den gleichen Raum Ω als Menge der Elementarereignisse und die gleiche σ-Algebra der Ereignisse $\mathcal{E}$ haben:

Definition 2.1.1. *Ein* statistisches Modell *besteht aus*

[1)] Und auch mathematisch zu rechtfertigen: Man muss nur zu geeigneten Produkten von Maßräumen übergehen.

(i) Einer Menge Ω und einer σ-Algebra $\mathcal{E}$ auf Ω.

(ii) Einer durch eine Menge Θ parametrisierten Familie $\mathbb{P}_\theta$ von Wahrscheinlichkeitsmaßen auf $(\Omega, \mathcal{E})$.

In Kurzfassung schreibt man dafür $(\Omega, \mathcal{E}, (\mathbb{P}_\theta)_{\theta\in\Theta})$.

Bemerkungen und Beispiele:

1. Betrachte etwa Bernoulliexperimente mit einer unbekannten Erfolgswahrscheinlichkeit p. Dann ist $\Omega = \{0, 1\}$, $\Theta = [\,0, 1\,]$, und für $p \in \Theta$ ist $\mathbb{P}_p$ durch $\mathbb{P}_p(\{1\}) := p$ definiert.

2. Werden normalverteilte Daten mit bekannter Streuung beobachtet, so wird Ω gleich $\mathbb{R}$ sein, versehen mit den Borelmengen als σ-Algebra, und Θ entspricht der Menge der möglichen Erwartungswerte.

3. Ω sei wie vorstehend, und $\mathbb{P}_a$ bezeichne für $a > 0$ die Gleichverteilung auf $[\,0, a\,]$. Dieses Beispiel wird gleich noch etwas eingehender studiert.

4. Das Ausgehen von einem statistischen Modell ist der erste Schritt zur mathematischen Behandlung eines Problems. Dieser Schritt fasst die über das Problem vorliegende Information zusammen: Mit welchen Zufallsautomaten ist überhaupt zu rechnen?

Es gibt aber eine Reihe von Schwierigkeiten: Warum ist die Einschränkung auf gerade diese Wahrscheinlichkeitsräume gerechtfertigt? Ist die Verteilungsannahme vielleicht nur approximativ richtig?

5. In der Regel ist – wie in den vorstehenden Beispielen – die Menge Ω eine Teilmenge von $\mathbb{R}$. Das ergibt sich auch in allgemeineren Situationen, wenn als Stichprobe reellwertige Zufallsvariable „abgefragt“ werden und man das induzierte Wahrscheinlichkeitsmaß betrachtet.

Und nun soll geschätzt werden. Etwas präziser soll das bedeuten, dass den Parametern $\theta \in \Theta$ eine Zahl $\gamma(\theta)$ zugeordnet ist, die uns aus irgendwelchen Gründen interessiert. Formal liegt also eine Abbildung $\gamma : \Theta \to \mathbb{R}$ vor.

> Manchmal ist die Bildmenge auch der $\mathbb{R}^k$, für unsere Zwecke reicht es aber, sich auf Skalare zu konzentrieren. Wir vereinbaren hier auch gleich für den Rest der Vorlesung: Ist Θ eine Teilmenge des $\mathbb{R}^m$, so soll γ messbar sein, wenn man in Bild- und Urbildbereich die Borelmengen betrachtet. *Überhaupt soll im Folgenden gelten:* Alles, was wir aufschreiben, soll definiert sein; es soll nicht jedesmal neu gesagt werden, dass irgendwelche Mengen oder Abbildungen messbar sein sollen.

Beispiele:

1. Mal angenommen, unser statistisches Modell besteht aus allen Normalverteilungen auf $\mathbb{R}$, die Parametermenge Θ ist also die Menge der (μ, σ) mit $\mu \in \mathbb{R}$

und $\sigma > 0$. Dann könnte $\gamma(\mu, \sigma) := \mu$ interessant sein: Welche Schätzung kann für den Erwartungswert abgegeben werden?

2. Das statistische Modell bestehe aus allen Poissonverteilungen auf $\mathbb{N}_0$, hier ist also Θ die Menge der $\lambda > 0$. Nun wird n-mal abgefragt, und aufgrund des Ausgangs sollen wir die Wahrscheinlichkeit dafür schätzen, dass zwei unabhängige weitere Abfragen beide zum Ergebnis 0 führen. Diese Wahrscheinlichkeit ist gleich $e^{-2\lambda}$, das folgt aus der Formel für die Poissonverteilung und der Unabhängigkeit. Folglich wäre hier $\gamma(\lambda) := e^{-2\lambda}$ zu betrachten.

3. In sehr vielen Fällen ist Θ eine Teilmenge der reellen Zahlen und γ die Identität. Da möchte man einfach wissen, mit welchem Wahrscheinlichkeitsmaß man es zu tun hatte.

2.2 Güteeigenschaften von Schätzern

Fassen wir das Bisherige zusammen, so haben wir ein statistisches Modell und eine Abbildung $\gamma : \Theta \to \mathbb{R}$ vorgelegt bekommen. Es stehen n unabhängige Abfragen aus dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{E}, \mathbb{P}_\theta)$ (mit einem unbekannten Parameter θ) zur Verfügung, und wir sollen daraus „möglichst gut“ den Wert $\gamma(\theta)$ schätzen.

Wir betrachten **zum Schätzproblem das folgende Beispiel**. In einem Quiz spielen zwei Kandidaten mit. Der Quizmaster sucht sich eine Zahl $a > 0$ und produziert dann n unabhängige Abfragen $x_1, \ldots, x_n$ aus der Gleichverteilung auf $[\,0, a\,]$. Diese Zahlen werden den Kandidaten mitgeteilt, und sie sollen daraus einen Wert für a schätzen. Wer näher dran ist, hat gewonnen.

Spieler A überlegt so: Der Erwartungswert der Gleichverteilung ist $a/2$, und für große n sollte das Stichprobenmittel nach den Gesetzen der großen Zahlen in der Nähe des Erwartungswertes liegen. Also schätzt er den Wert

$$\frac{2}{n}(x_1 + \cdots + x_n).$$

Spieler B argumentiert anders, er analysiert das Verhalten von $\max\{X_1, \ldots, X_n\}$ von unabhängigen Zufallsvariablen X_i, die alle auf $[\,0, a\,]$ gleichverteilt sind. Die Wahrscheinlichkeit, dass alle $\leq a - \varepsilon$ sind, ist doch $\left(\dfrac{a-\varepsilon}{a}\right)^n$, und das geht für $n \to \infty$ gegen Null. Folglich wird der Wert

$$\max\{x_1, \ldots, x_n\}$$

mit hoher Wahrscheinlichkeit nahe bei a liegen, *das* ist die von B favorisierte Schätzung.

Es gibt also mehrere naheliegende Schätzer, und es ist nicht klar, welche Methode vorzuziehen ist.

Am vorstehenden Beispiel kann ein wichtiger Begriff erläutert werden: Wenn durch irgendeine Formel aus den Messwerten eine Zahl entsteht, so soll das ein *Schätzer für* γ heißen. Eben hatten wir das doppelte Stichprobenmittel und das Maximum betrachtet, die allgemeine Definition ist wie folgt:

Definition 2.2.1. *Ein statistisches Modell und eine Funktion* γ *seien wie vorstehend gegeben. Unter einem* Schätzer für γ auf der Grundlage von n Messungen *verstehen wir dann eine messbare Abbildung*

$$d : \Omega^n \to \mathbb{R}.$$

Bemerkungen:

1. Die Interpretation ist die folgende: Ist $\mathbb{P}_\theta$ das „richtige“ Wahrscheinlichkeitsmaß und ergibt das Experiment die Werte $x_1, \ldots, x_n$, so soll als Vorschlag für die Schätzung von $\gamma(\theta)$ die Zahl $d(x_1, \ldots, x_n)$ angegeben werden.

2. Die Definition ist sehr allgemein, es sind im Prinzip auch noch völlig sinnlose Funktionen als „Schätzer“ möglich.

3. Ideal wäre natürlich ein Schätzer, der den richtigen Wert immer genau trifft: Egal, was θ ist und wie die $x_1, \ldots, x_n$ ausfallen, es ist $d(x_1, \ldots, x_n) = \gamma(\theta)$.

   So etwas kann vorkommen: Wenn das statistische Modell aus allen Punktmaßen δ_a auf $\mathbb{R}$ besteht (es wird also mit Wahrscheinlichkeit 1 die Zahl a erzeugt) und a zu schätzen ist, so kann man

   $$d(x_1, \ldots, x_n) := x_1$$

   wählen. Leider ist nur in solch extremen Fällen zu erwarten, dass man das Schätzproblem so vollständig lösen kann.

Im Allgemeinen wird man mit weniger zufrieden sein müssen, doch was sollen die wichtigen Forderungen sein? Im Rest dieses Abschnitts beschäftigen wir uns noch mit einigen Kriterien, um die Güte von Schätzern zu messen.

Eine plausible Forderung ist sicher, dass ein Schätzer *erwartungstreu* ist. Das soll folgendes bedeuten. Mal angenommen, es geht um ein spezielles θ, *dieser* Zufallsautomat wurde gewählt. Der wird nun n-mal abgefragt, und wir geben die Prognose $d(x_1, \ldots, x_n)$. Wenn wir das sehr oft machen, so wird sich im Mittel der Erwartungswert von $d : \Omega^n \to \mathbb{R}$ ergeben; dabei trägt Ω^n das n-fache Produktmaß zu $\mathbb{P}_\theta$, und wir nehmen an, dass der Erwartungswert existiert. Ein Schätzer sollte mindestens im Mittel richtige Werte liefern, und das führt zu

Definition 2.2.2. *Ein Schätzer heißt* erwartungstreu*, wenn für alle* θ *der Erwartungswert von* d *gleich* $\gamma(\theta)$ *ist.*

Bemerkungen:

1. Ist d die konstante Abbildung $\gamma(\theta_0)$, so liefert sie zwar für den Spezialfall $\theta = \theta_0$ hervorragende Ergebnisse, sie versagt aber für andere θ und ist insbesondere nicht erwartungstreu.

2. Wir betrachten ein aus zwei Karten bestehendes Kartenspiel, aus dem eine einzelne Karte gezogen wird. Wir ziehen n mal mit Zurücklegen und sollen daraufhin die Anzahl θ der roten Karten in dem Spiel schätzen. Es ist also $\Theta = \{0, 1, 2\}$, und γ ist die identische Abbildung. Ein plausibler, sogar erwartungstreuer, Schätzer ist „zwei mal Anzahl der roten Karten in der Stichprobe, geteilt durch n“, das folgt daraus, dass der Erwartungswert der Binomialverteilung gleich np ist. Das führt allerdings zu der merkwürdigen Situation, dass man eventuell Schätzungen der Form „Das Spiel enthält 0.92 rote Karten“ abgeben muss.
Übrigens: Logischer wäre hier, nur Schätzer mit Werten in $\{0, 1, 2\}$ zuzulassen.

Übung: Gibt es für jedes n erwartungstreue Schätzer dieser Art?

Wir wollen die Strategien von Spieler A und B analysieren. Für Spieler A ist das einfach: Der Erwartungswert jeder Einzelabfrage ist $a/2$, folglich hat auch der Mittelwert aus n Abfragen diesen Erwartungswert und damit ist die Schätzung von A – der doppelte Mittelwert – erwartungstreu. Für die Diskussion der Strategie von Spieler B ist an einige Sachverhalte aus der elementaren Stochastik zu erinnern:

- Sind $X_1, \ldots, X_n$ unabhängig, so ist
$$\mathbb{P}(\max\{X_1, \ldots, X_n\} \leq c) = \prod_i \mathbb{P}\{X_i \leq c\}.$$
- Sind insbesondere die X_i unabhängige Abfragen der Gleichverteilung auf $[\,0, a\,]$, so ist die Wahrscheinlichkeit, dass das Maximum $\leq c$ ist, durch $(c/a)^n$ gegeben.
- Ist μ ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ mit einer Dichte f und kennt man die Funktion $x \mapsto \mu(]\,-\infty, x\,])$, so ist f die Ableitung dieser Funktion.
- Im vorliegenden Fall heißt das: Das Maximum aus n Gleichverteilungen auf $[\,0, a\,]$ hat eine Dichtefunktion, nämlich die Funktion nx^{n-1}/a^n (definiert auf $[\,0, a\,]$).
- Hat ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ eine Dichte f, so ist der Erwartungswert der Identität durch $\int xf(x)dx$ gegeben.
- Es folgt: Das Maximum aus n Gleichverteilungen auf $[\,0, a\,]$ hat den Erwartungswert
$$\int_0^a nxx^{n-1}/a^n dx = \frac{n}{n+1}a.$$

Damit sieht man, dass B nicht erwartungstreu schätzt. Das ist aber leicht zu reparieren, wir empfehlen B, seinen Wert mit $(n + 1)/n$ zu multipizieren, also als Schätzung
$$\frac{n+1}{n} \max\{x_1, \ldots, x_n\}$$

zu wählen. *Dann* hat auch B einen erwartungstreuen Schätzer.

Ein weiteres Kriterium für die Güte eines Schätzers ist sicherlich, wie stark die Schätzwerte um den wirklichen Wert schwanken. Im Idealfall sollte nicht nur der Erwartungswert von d unter $\mathbb{P}_\theta^n$ gleich $\gamma(\theta)$ sein, schön wäre auch, wenn die Streuung klein wäre, dass also garantiert werden kann, dass die Irrtümer kontrollierbar bleiben.

Wie sieht es denn bei Spieler A und Spieler B aus? Wieder ist Spieler A einfacher zu behandeln: Die Varianz der Gleichverteilung auf $[\,0, a\,]$ ist bekanntlich gleich $a^2/12$. Der Mittelwert aus n Abfragen führt zu einem n im Nenner, und da A den Mittelwert mit 2 multipliziert hat, erhalten wir noch einen Faktor 4. Fasst man alles zusammen, so ergibt sich für die Varianz des A-Schätzers der Wert $\dfrac{a^2}{3n}$.

Nun zu B. Wir nutzen aus, dass der Erwartungswert des Quadrats der Identität das Integral über x^2 mal Dichtefunktion ist und dass die Gleichung

$$Var(X) = E(X^2) - (E(X))^2$$

gilt. So folgt: Die Varianz des Maximums aus n Gleichverteilungen ist gleich

$$\int_0^a n\, x^2 x^{n-1} a^{-n} dx - \left(\frac{na}{n+1}\right)^2 = \frac{na^2}{(n+1)^2(n+2)}.$$

Damit gilt: Der B-Schätzer hat eine wesentlich bessere Varianz, allerdings schwankt er um den falschen Wert. Macht man noch die $(n+1)/n$-Modifikation, so ist die eben berechnete Zahl mit dem Quadrat von $(n+1)/n$ zu multiplizieren, es ergibt sich die Varianz

$$\frac{(n+1)^2}{n^2}\frac{na^2}{(n+1)^2(n+2)} = \frac{a^2}{n(n+2)}.$$

Was wäre, wenn B nicht modifiziert, wie groß wäre dann die mittlere quadratische Abweichung? Wir rechnen so:

$$\begin{aligned}
E_{\mathbb{P}_a}\left(\left(\max\{X_1,\ldots,X_n\} - a\right)^2\right) &= E_{\mathbb{P}_a}\left(\left(\max\{X_1,\ldots,X_n\} - \frac{n}{n+1}a - \frac{1}{n+1}a\right)^2\right) \\
&= E_{\mathbb{P}_a}\left(\left(\max\{X_1,\ldots,X_n\} - \frac{n}{n+1}a\right)^2\right) + \frac{a^2}{(n+1)^2} \\
&= Var_{\mathbb{P}_a}(\max\{X_1,\ldots,X_n\}) + \frac{a^2}{(n+1)^2} \\
&= \frac{na^2}{(n+1)^2(n+2)} + \frac{a^2}{(n+1)^2} \\
&= \frac{2a^2}{(n+1)(n+2)}.
\end{aligned}$$

Zusammen: B ist im Mittel auch dann wesentlich besser als A, wenn nicht modifiziert wurde und n groß ist. Doch hat B schon die bestmögliche Strategie? Es soll in den folgenden Abschnitten versucht werden, diese Frage sinnvoll zu formulieren und in Spezialfällen eine Lösung zu finden.

Vorher behandeln wir noch einige *Ergänzungen zu erwartungstreuen Schätzern.* Als erstes diskutieren wir ein *Gegenbeispiel*: Manchmal gibt es überhaupt keine erwartungstreuen Schätzer:

> Das statistische Modell bestehe aus allen möglichen hypergeometrischen Verteilungen zu festem r auf $\{0,\dots,m\}$: Es gibt also Urnen mit n Kugeln (wobei $n \geq r$), von denen r rot und $n-r$ weiß sind, und aus denen wird m-mal ohne Zurücklegen gezogen; dabei läuft n durch alle natürlichen Zahlen $\geq r$. Wir fragen uns, ob es einen erwartungstreuen Schätzer für n gibt[2)]. Sei nun d irgendein Schätzer, der aus der Anzahl der konkret gezogenen roten Kugeln ein N schätzt. Wegen der Endlichkeit der Menge $\{0,\dots,m\}$ ist d beschränkt. Andererseits können die n-Werte beliebig groß werden, und deswegen werden sie bestimmt nicht alle als Erwartungswert von d auftreten.

Es folgen nun zwei wichtige Beispiele für erwartungstreue Schätzer:

Satz 2.2.3. *Gegeben sei ein beliebiges statistisches Modell.*

(i) Das Stichprobenmittel ist ein erwartungstreuer Schätzer für den Erwartungswert der Identität auf Ω.

(ii) Die Stichprobenvarianz schätzt die Varianz erwartungstreu. Deswegen *wird im Nenner $n-1$ gewählt, die Zahl n hätte nicht zu einem erwartungstreuen Schätzer geführt.*

Beweis: (i) Das ist klar, man muss nur wissen, dass der Erwartungswert eine lineare Funktion ist.

(ii) Die Behauptung ist eine Umformulierung der folgenden Aussage:
Sind $X_1,\dots,X_n$ unabhängige Kopien einer Zufallsvariable X und bezeichnet man mit $\overline{X}$ die Zufallsvariable $(X_1+\cdots+X_n)/n$, so gilt

$$E\left(\frac{1}{n-1}\sum_i (X_i-\overline{X})^2\right) = \sigma^2(X).$$

Zur Vorbereitung des Beweises bemerken wir, dass

$$\begin{aligned} E[(X_i-E(X))(\overline{X}-E(X))] &= \sigma^2/n, \\ E(\overline{X}-E(X))^2 &= \sigma^2/n; \end{aligned}$$

beides liegt an der Definition von σ^2 und der Tatsache, dass der Erwartungswert für unabhängige Zufallsvariable multiplikativ ist.

[2)] Man sollte sich hier an die Illustration dieses Problems durch die Fische-Anzahl aus der elementaren Stochastik erinnern.

Nun können wir so rechnen:

$$\begin{aligned} E\left(\frac{1}{n-1}\sum_i (X_i - \overline{X})^2\right) &= E\left(\frac{1}{n-1}\sum_i [(X_i - E(X)) - (\overline{X} - E(X))]^2\right) \\ &= E\Big(\frac{1}{n-1}\sum_i [(X_i - E(X))^2 + \\ &\quad -2(X_i - E(X))(\overline{X} - E(X)) + (\overline{X} - E(X))^2]\Big) \\ &= \frac{1}{n-1}\sum_i \big[\sigma^2 - \frac{2}{n}\sigma^2 + \frac{1}{n}\sigma^2\big] \\ &= \sigma^2. \end{aligned}$$

□

Wir behandeln nun einen weiteren wichtigen Begriff. Es soll ausgedrückt werden, dass eine Folge von Schätzern den Zielwert mit Sicherheit besser und besser approximiert. Hier die exakte Formulierung:

Definition 2.2.4. *Ein statistisches Modell und eine Funktion γ seien wie vorstehend gegeben. Weiter sei für jedes n ein Schätzer definiert, der aus n Abfragen einen Schätzwert für $\gamma(\theta)$ erzeugt. Diese Folge (d_n) von Schätzern für $\gamma(\theta)$ heißt eine* konsistente Schätzfolge für γ*, wenn die (d_n) für jedes $\mathbb{P}_\theta$ in Wahrscheinlichkeit gegen $\gamma(\theta)$ konvergieren. Genauer: Für jedes $\varepsilon > 0$ soll*

$$\mathbb{P}_\theta(\{(x_1, \ldots, x_n) \in \Omega^n \mid |d_n(x_1, \ldots, x_n) - \gamma(\theta)| > \varepsilon\})$$

für $n \to \infty$ gegen Null gehen.

Bemerkungen und Beispiele:

1. Die Definition ist von eher theoretischem Interesse, sie liefert nur einen groben Gütemaßstab für Schätzer.

2. Das schwache Gesetz der großen Zahlen besagt, dass die Folge der Stichprobenmittel eine konsistente Schätzfolge für den Erwartungswert darstellt.

Ab jetzt behandeln wir nur noch erwartungstreue Schätzer für ein vorgelegtes statistisches Modell und eine fest vorgegebene Funktion γ. Ziel ist doch, γ möglichst genau zu schätzen, d.h. beim Schätzen möglichst kleine Varianzen zu erreichen. Das legt die folgende Definition nahe:

Definition 2.2.5. *Ein erwartungstreuer Schätzer d^* für γ heißt* gleichmäßig bester erwartungstreuer Schätzer*, falls gilt: Ist d ebenfalls ein erwartungstreuer Schätzer, so ist für jedes θ*

$$Var_{\mathbb{P}_\theta}(d) \geq Var_{\mathbb{P}_\theta}(d^*).$$

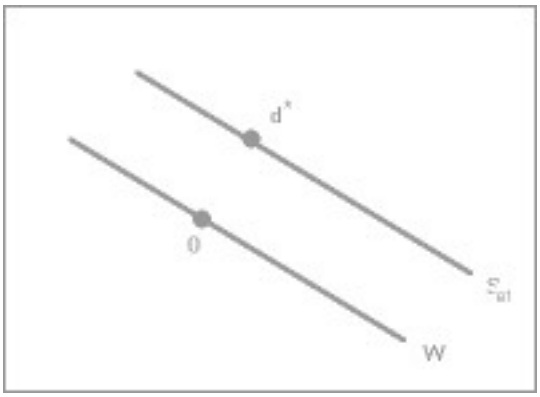
Bemerkung: Es ist überhaupt nicht klar, dass es so etwas gibt[3)], und naheliegende Kandidaten sind evtl. nicht bestmöglich. Als Beispiel betrachten wir die Situation, wo $\Omega = \{1, 2, 3, 4, 5, 6\}$ ist und Θ zweielementig ist: Zur Wahl stehen nur die Gleichverteilungen auf $\{1, 2, 3\}$ und $\{4, 5, 6\}$, und zu schätzen ist der Erwartungswert aufgrund einer Messung. Dann ist das Stichprobenmittel (das ist hier die Identität) zwar erwartungstreu, es geht aber offensichtlich besser. Ein Schätzer, der auf $\{1, 2, 3\}$ den Wert 2 und auf dem Rest den Wert 5 annimmt, schätzt mit Varianz Null. Er ist damit sicher ein gleichmäßig bester Schätzer, doch wie findet man so etwas in schwierigeren Fällen?

Wir wollen noch versuchen, uns optimale erwartungstreue Schätzer *geometrisch* zu veranschaulichen. Dazu fixieren wir zunächst ein $\mathbb{P}_\theta$ und betrachten den Raum V aller Funktionen auf dem $\mathbb{R}^n$, die bezüglich des n-fachen Produktmaßes quadrat-integrabel sind. *Das* sind die hier interessierenden Schätzfunktionen d. Der Unterraum W aller Funktionen mit Erwartungswert 0 spielt eine wichtige Rolle: Ist nämlich d ein erwartungstreuer Schätzer für $\gamma(\theta)$, so besteht die Menge S_{et} *aller* erwartungstreuen Schätzer aus den Funktionen $d + h$ mit $h \in W$, es handelt sich also um einen affinen Unterraum.

Nun ist für jedes f die Varianz gleich

$$\langle f, f\rangle - (E(f))^2,$$

und da alle $f \in S$ den gleichen Erwartungswert haben, heißt das: Ein erwartungstreuer Schätzer hat genau dann die kleinste Varianz, wenn es sich um ein Element aus S_{et} mit kleinstem Abstand zum Nullpunkt handelt. Außerdem sieht man sofort, dass es höchstens einen solchen besten Schätzer geben kann und dass er dadurch charakterisiert ist, dass er auf allen Elementen aus W senkrecht steht.



Wenn man nun noch *alle* $\mathbb{P}_\theta$ berücksichtigt, muss man W durch den Durchschnitt der zu den einzelnen θ gehörigen Räumen ersetzen. Eindeutigkeit und Orthogonalitäts-Charakterisierung bleiben erhalten, und so ergibt sich ein bekanntes klassisches Kriterium von Rao.

[3)] Sicher ist eine notwendige Voraussetzung, dass es überhaupt erwartungstreue Schätzer gibt.

2.3 Suffizienz und Vollständigkeit: diskrete Räume

Wie findet man nun aber beste Schätzer? Wir behandeln dazu ein weiteres wichtiges Konzept, mit dem wir schließlich in der Lage sein werden, in vielen wichtigen Fällen beste Schätzer konkret anzugeben. Es geht im Wesentlichen um *Informationsreduktion*, intuitiv ist es leicht zu verstehen: Welche Informationen in der Stichprobe könnten für die Berechnung der Schätzfunktion an dieser Stelle wirklich eine Rolle spielen? Ein Beispiel: Soll der Anteil der roten Karten in einem Kartenspiel geschätzt werden und ergibt sich bei unabhängiger Ziehung (mit Zurücklegen) das Ergebnis $r, s, r, r, s, s, r, r, r, r$, so ist doch plausibel, dass die Reihenfolge der gezogenen Farben sicher keine Rolle spielt und dass es sogar nur auf die Anzahl der in der Stichprobe gefundenen roten Karten ankommt.

Das ist nicht so einfach zu präzisieren. Wir betrachten unser übliches statistisches Modell: Es ist $\Omega \subset \mathbb{R}$, und eine Familie $(\mathbb{P}_{\theta\in\Theta})$ von Wahrscheinlichkeitsmaßen ist gegeben. Nun wird n-mal abgefragt, wobei θ unbekannt ist. Wir erhalten $x_1, \ldots, x_n$ und versuchen nun, die darin enthaltenen Informationen zu komprimieren. Es soll alles weggelassen werden, was wir sowieso schon wissen.

> **Um die zu besprechenden Ideen nicht durch neue technische Definitionen zu verschleiern, werden wir im Folgenden meist annehmen, dass Ω endlich oder höchstens abzählbar ist.**

Hier ist die entscheidende

Definition 2.3.1. *Sei $T : \Omega^n \to \mathbb{R}^m$ eine messbare Abbildung. T heißt* suffizient *für $(\mathbb{P}_{\theta\in\Theta})$, wenn gilt: Ist y im Bild von T, so sind die Einschränkungen aller $\mathbb{P}_\theta^n$ auf*

$$\Omega_y := \{x \mid Tx = y\} \subset \Omega^n$$

identisch. Anders ausgedrückt: Es gibt ein Wahrscheinlichkeitsmaß Q_y auf Ω_y, so dass

$$Q_y = (\mathbb{P}_\theta^n)|_{\Omega_y}$$

für alle θ [4)].

Das ist, zugegeben, schwierig, und deswegen gibt es einige

Bemerkungen und Beispiele:

1. T bewirkt so etwas wie eine Zerlegung von Ω^n, und genau genommen ist nur diese Zerlegung interessant[5)].

2. Es sei $\Omega = \{0, 1\}$, das statistische Modell bestehe aus allen Bernoulliverteilungen zu $p \in [\,0, 1\,]$; dann sind die $\mathbb{P}_p^n$ die Produktmaße auf $\{0, 1\}^n$. Wir wollen die

[4)] Die Einschränkung eines Maßes $\mathbb{P}$ ist durch $\mathbb{P}|_B(E) := \mathbb{P}(E)/\mathbb{P}(B)$ für $E \subset B$ erklärt. Sie ist offensichtlich nur für $\mathbb{P}(B) > 0$ definiert, im Folgenden müsste es also hin und wieder den Zusatz „falls definiert“ geben. Alternativ: Man kann sich erst einmal auf den Fall beschränken, dass stets $\mathbb{P}_\theta(\Omega_y) > 0$ gilt.

[5)] Das passt natürlich in das Konzept der Wahrscheinlichkeitstheorie, dass Information in Unter-σ-Algebren verschlüsselt ist.

Abbildung $T : (x_1, \ldots, x_n) \mapsto \sum_i x_i$ untersuchen, wir behaupten, dass sie suffizient ist. Sei k fest (das entspricht dem y der allgemeinen Definition), wir wollen uns auf die Menge aller $(x_1, \ldots, x_n)$ konzentrieren, die genau k Einsen enthalten. Bei vorgelegtem $\mathbb{P}_p$ ist die Wahrscheinlichkeit dieser Menge gleich $b(k, n; p)$, und jedes einzelne Element hat die gleiche Wahrscheinlichkeit (nämlich den $\binom{n}{k}$-ten Anteil davon). Folglich kann man Q_k als die Gleichverteilung auf Ω_k wählen, und die Suffizienz ist bewiesen.

> Eine Variante: Teilt man $\{1, \ldots, n\}$ in s disjunkte nicht leere Teilmengen auf, so ist der aus den Summen über die jeweiligen Anteile bestehende Vektor (des $\mathbb{R}^s$) eine suffiziente Statistik.

3. Betrachte das gleiche Modell wie eben, diesmal aber $T : (x_1, \ldots, x_n) \mapsto x_1$. Diese Abbildung schaut sich also nur das erste Ergebnis an. Ein möglicher T-Wert ist $k = 1$, die Menge Ω_1 besteht aus allen n-Tupeln in Ω^n, die an der ersten Stelle eine 1 enthalten. Die Wahrscheinlichkeit unter $\mathbb{P}_p^n$ ist p, die bedingte Wahrscheinlichkeit der Menge der Teilmenge $(1, x_2, \ldots, x_n)$, die außer der ersten Eins k' Einsen enthalten, ist damit gleich $b(k', n-1, p)$ (dabei sei $0 \leq k' \leq n-1$ beliebig). Und diese Zahlen sind *nicht* unabhängig von p, Suffizienz liegt also nicht vor.

4. Ein triviales Beispiel: Die Identität ist immer suffizient, allgemeiner jede injektive Abbildung. Umgekehrt kann es vorkommen, dass es nur injektive suffiziente Abbildungen gibt, d.h., dass keine Informationsreduktion möglich ist.

> Betrachte ein dreielementiges Ω und wähle darauf so viele Wahrscheinlichkeitsmaße (d.h., Wahrscheinlichkeitsvektoren), dass sie auf keiner zweielementigen Teilmenge proportional sind. Dann müssen die $T^{-1}(y)$ für suffiziente Abbildungen einelementig sein.
>
> (Noch einfacher: Auf einem zweielementigen Ω reichen sogar zwei Maße.)

Das *weitere Vorgehen* ist wie folgt:

- Bevor es richtig losgeht, benötigen wir einige *Vorbereitungen*.
- Zunächst zeigen wir, wie sich aus einem erwartungstreuen Schätzer und einer suffizienten Statistik[6)] auf naheliegende Weise ein weiterer Schätzer konstruieren lässt, der für jedes θ zu einer nicht schlechteren Varianz führt und nur noch von $T(x)$ abhängt.
- Damit sollte klar sein, dass es sich um einen hier wichtigen Begriff handelt. Wir diskutieren ein Kriterium für Suffizienz und behandeln dann weitere Beispiele.

[6)] „Statistik“ wird in diesem Zusammenhang synonym mit „Abbildung“ verwendet. Diese Bezeichnung ist historisch bedingt.

- Dann muss noch ein technischer Begriff eingeführt werden: *Vollständigkeit.* Unter Verwendung dieser Definition ist es möglich, den Hauptsatz der Theorie, den Satz von Lehmann-Scheffé, zu formulieren und zu beweisen.
- Es folgen noch einige Beispiele, um mit diesem Satz bestmögliche Schätzer konkret zu konstruieren.

Vorbereitungen

1. Sei K eine Teilmenge eines $\mathbb{R}$-Vektorraumes. K heißt *konvex*, wenn K alle Konvexkombinationen von Elementen aus K enthält: Ist $n \in \mathbb{N}$ und sind $x_1, \ldots, x_n \in K$, $\lambda_1, \ldots, \lambda_n \geq 0$ mit $\sum_i \lambda_i = 1$, so liegt die Konvexkombination $\sum_i \lambda_i x_i$ ebenfalls in K.

2. Sei K konvex und $f : K \to \mathbb{R}$ eine Abbildung. f heißt *konvex*, wenn für alle Konvexkombinationen gilt:

$$f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i).$$

Gilt stets $\geq$ bzw. stets $=$, so heißt f *konkav* bzw. *affin*.

> Es reicht, die Ungleichung für Konvexkombinationen aus zwei Elementen zu fordern (einfache Übungsaufgabe).
>
> Beispiele: Wenn $K = \mathbb{R}$ ist, so sind die affinen Abbildungen genau die Abbildungen der Form $x \mapsto ax+b$; für zweimal differenzierbare Funktionen ist $f'' \geq 0$ notwendig und hinreichend für Konvexität. Auf dem $\mathbb{R}^n$ kann man hinreichende Bedingungen mit Hilfe der Hesse-Matrix formulieren.

3. Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable und $f : \mathbb{R} \to \mathbb{R}$ eine konvexe Funktion. Dann gilt

$$f(E(X)) \leq E(f(X));$$

das ist die *Jensensche Ungleichung.*
(Beweis: Wir betrachten nur den Fall endlicher Räume Ω. Da folgt das Ergebnis direkt aus der Konvexitätsdefinition, denn der Erwartungswert von X ist die Konvexkombination mit Gewichten $\mathbb{P}(\{\omega\})$ der $X(\omega)$.)

4. Der diskrete Wahrscheinlichkeitsraum Ω sei disjunkt in $\Delta_1 \cup \cdots \cup \Delta_r$ zerlegt, und $X : \Omega \to \mathbb{R}$ sei eine Zufallsvariable. Unter dem *bedingten Erwartungswert von X unter Δ_j* versteht man dann den Erwartungswert der Einschränkung von X auf den Wahrscheinlichkeitsraum Δ_j, also die Zahl

$$\frac{\sum_{x \in \Delta_j} X(x)\mathbb{P}(\{x\})}{P(\Delta_j)}.$$

(Inhaltliche Interpretation: Mit was für einem X-Wert sollte man im Mittel rechnen, wenn man weiß, dass bestimmt ein Ergebnis in Δ_j herauskommt. Besteht die Zerlegung nur aus einem Element, kommt der gewöhnliche Erwartungswert heraus. Wir werden das Thema in Abschnitt 4.5 noch einnmal aufgreifen.)

Suffizienz verbessert Schätzer

Wir kümmern uns nun wieder um das Schätzproblem, gehen also wieder von einem statistischen Modell (mit diskretem Ω) und einer bekannten Zielfunktion $\gamma : \Theta \to \mathbb{R}$ aus. Weiter sollen eine suffiziente Statistik $T : \Omega^n \to \mathbb{R}^m$ und ein erwartungstreuer Schätzer $d : \Omega^n \to \mathbb{R}$ für γ vorgegeben sein.

Definition 2.3.2. *Sei y im Bild von T. Nach Voraussetzung gibt es ein Wahrscheinlichkeitsmaß Q_y auf $\Omega_y := T^{-1}(y)$, so dass alle durch die $\mathbb{P}_\theta$ auf Ω_y induzierten Wahrscheinlichkeitsmaße mit Q_y übereinstimmen (falls sie definiert sind).*

Mit $d_T(y)$ bezeichnen wir den Erwartungswert der Einschränkung von d auf Ω_y unter Q_y.

(Diese Zahl ist also das, was man als d-Wert unter $\mathbb{P}_\theta^n$ erwarten würde, wenn man weiß, dass $T = y$ ist; nach Voraussetzung ist diese Zahl unabhängig vom Parameter θ.)

In Formeln:

$$d_T(y) := \frac{\sum_{x \in \Omega^n, Tx=y} d(x)\mathbb{P}_\theta^n(\{x\})}{\mathbb{P}_\theta^n(\Omega_y)},$$

und dieser Ausdruck ist unabhängig von θ. Auf diese Weise wird eine Abbildung d_T vom Bild von T nach $\mathbb{R}$ induziert.

Der Schätzer $d_T \circ T$ ist nicht schlechter als d:

Satz 2.3.3. *Für den vorstehend definierten Schätzer $d_T \circ T$ gilt:*

(i) $d_T \circ T$ ist erwartungstreu.

(ii) Die Varianzen sind nicht schlechter als die von d: Für alle θ ist

$$Var_\theta(d) \geq Var_\theta(d_T \circ T).$$

Dabei steht Var_θ für die Varianz unter Verwendung von $\mathbb{P}_\theta^n$ auf Ω^n, entsprechend ist der Erwartungswert E_θ zu interpretieren.

Beweis: (i) Das ergibt sich wie folgt aus der Definition:

$$\begin{aligned}
E_\theta(d_T \circ T) &= \sum_{x\in\Omega^n} (d_T \circ T)(x)\mathbb{P}_\theta^n(\{x\}) \\
&= \sum_{y\in T(\Omega^n)} \sum_{x\in\Omega_y} (d_T \circ T)(x)\mathbb{P}_\theta^n(\{x\}) \\
&= \sum_{y\in T(\Omega^n)} \sum_{x\in\Omega_y} d_T(y)\mathbb{P}_\theta^n(\{x\}) \\
&= \sum_{y\in T(\Omega^n)} d_T(y) \sum_{x\in\Omega_y} \mathbb{P}_\theta^n(\{x\}) \\
&= \sum_{y\in T(\Omega^n)} d_T(y)\mathbb{P}_\theta^n(\Omega_y) \\
&= \sum_{y\in T(\Omega^n)} \frac{\sum_{x\in\Omega_y} d(x)\mathbb{P}_\theta^n(\{x\})}{\mathbb{P}_\theta^n(\Omega_y)} \mathbb{P}_\theta^n(\Omega_y) \\
&= \sum_x d(x)\mathbb{P}_\theta^n(\{x\}) \\
&= E_\theta(d) \\
&= \gamma(\theta).
\end{aligned}$$

(ii) Hier geht an entscheidender Stelle die Jensensche Ungleichung ein. Wir fixieren θ und betrachten die disjunkte Zerlegung von Ω^n in die Ω_y, wobei y alle möglichen Bildwerte von T durchläuft.

Wir schauen uns für irgendein y das auf Ω_y durch $\mathbb{P}_\theta$ induzierte Wahrscheinlichkeitsmaß und die Zufallsvariable „Einschränkung von d auf Ω_y“ etwas genauer an. Da die Funktion $a \to (a-\alpha)^2$ für jedes α konvex ist, können wir $\alpha := \gamma(\theta)$ setzen und aus der Jensenschen Ungleichung folgern:

$$\sum_{x\in\Omega_y} (d(x)-\gamma(\theta))^2 \frac{\mathbb{P}_\theta^n(\{x\})}{\mathbb{P}_\theta^n(\Omega_y)} \geq \left(\sum_{x\in\Omega_y} d(x) \frac{\mathbb{P}_\theta^n(\{x\})}{\mathbb{P}_\theta^n(\Omega_y)} - \gamma(\theta) \right)^2 . \tag{2.1}$$

Für die Varianzen folgt damit die gewünschte Ungleichung:

$$\begin{aligned}
Var_\theta(d) &= \sum_{x\in\Omega^n} (d(x)-\gamma(\theta))^2 \mathbb{P}_\theta^n(\{x\}) \\
&= \sum_{y\in \mathrm{Bild}\, T} \sum_{x\in\Omega_y} (d(x)-\gamma(\theta))^2 \mathbb{P}_\theta^n(\{x\}) \\
&= \sum_{y\in \mathrm{Bild}\, T} \frac{\sum_{x\in\Omega_y}(d(x)-\gamma(\theta))^2 \mathbb{P}_\theta^n(\{x\})}{\mathbb{P}_\theta^n(\Omega_y)} \mathbb{P}_\theta^n(\Omega_y) \\
&= \sum_{y\in \mathrm{Bild}\, T} \left(\sum_{x\in\Omega_y} (d(x)-\gamma(\theta))^2 \frac{\mathbb{P}_\theta^n(\{x\})}{\mathbb{P}_\theta^n(\Omega_y)} \right) \mathbb{P}_\theta^n(\Omega_y) \\
&\geq \sum_{y\in \mathrm{Bild}\, T} \left(\sum_{x\in\Omega_y} d(x) \frac{\mathbb{P}_\theta^n(\{x\})}{\mathbb{P}_\theta^n(\Omega_y)} - \gamma(\theta) \right)^2 \mathbb{P}_\theta^n(\Omega_y) \\
&= \sum_{y\in \mathrm{Bild}\, T} (d_T(y)-\gamma(\theta))^2 \mathbb{P}_\theta^n(\Omega_y) \\
&= \sum_{x\in\Omega^n} (d_T\circ T(x)-\gamma(\theta))^2 \mathbb{P}_\theta^n(\{x\}) \\
&= Var_\theta(d_T\circ T).
\end{aligned}$$

□

Ein Kriterium für Suffizienz

Es ist günstig, im hier behandelten diskreten Fall ein einfaches Kriterium zur Verfügung zu haben. Es sei – unter den üblichen Voraussetzungen – eine Abbildung $T : \Omega^n \to \mathbb{R}^m$ vorgegeben. Suffizienz heißt dann doch, dass die Wahrscheinlichkeiten $\mathbb{P}_\theta$ auf den Ω_y proportional sind. Das führt unmittelbar zum nächsten Satz, dem so genannten *Neyman-Kriterium*[7]:

Satz 2.3.4. *T ist genau dann suffizient, wenn es eine Abbildung $h : \Omega^n \to \mathbb{R}$ und Abbildungen $g_\theta : \mathbb{R}^m \to \mathbb{R}$ (für $\theta \in \Theta$) so gibt, dass*

$$\mathbb{P}_\theta^n(\{x\}) = g_\theta(Tx)h(x)$$

für alle $x \in \Omega$.

Bemerkungen und Beispiele:

1. Ist T eine beliebige Abbildung und sind h und die g_θ ebenfalls beliebig mit positiven Werten vorgegeben, so kann man die Formel zur Definition von $\mathbb{P}_\theta$ verwenden; evtl. muss man g_θ noch mit einem Faktor multiplizieren, um zu Wahrscheinlichkeitsmaßen zu kommen. Anders ausgedrückt: *Alle* Abbildungen T können unter geeigneten Umständen als suffiziente Abbildungen vorkommen.

[7] Der Beweis ist leicht, er wird hier übersprungen.

2. Nach dem Satz liegt genau dann eine suffiziente Abbildung vor, wenn es gelingt, die Zahl $\mathbb{P}_\theta^n(\{x\})$ als „Funktion von x“ mal „(von θ möglicherweise abhängige) Funktion von Tx“ zu schreiben. Mit dieser Bemerkung lassen sich zahlreiche Beispiele finden, wir behandeln im Folgenden einige typische Vertreter.

3. Ein Bernoulli-0-1-Experiment werde n mal unabhängig wiederholt, die Erfolgswahrscheinlichkeit p sei unbekannt[8)]. Für ein spezielles $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ ist die Wahrscheinlichkeit unter $\mathbb{P}_p^n$ gleich „p hoch Anzahl der Einsen mal $1-p$ hoch Anzahl der Nullen in x“.
Setzt man $Tx := \sum_i x_i$, so ist

$$\mathbb{P}_\theta^n(\{x\}) \quad = \quad p^{Tx}(1-p)^{n-Tx}.$$

Mit $h = 1$ und $g_p(y) := p^y(1-p)^{n-y}$ ist damit das Neyman-Kriterium erfüllt, folglich ist T suffizient.

4. Wir betrachten $\Omega = \mathbb{N}_0$ und darauf alle Poisson-Verteilungen. Die Wahrscheinlichkeit für eine n-elementige Stichprobe unter einer speziellen, zum Parameter λ gehörigen Verteilung ist gleich

$$\mathbb{P}_\lambda^n(\{(x_1, \ldots, x_n)\}) \quad = \quad \frac{\lambda^{\sum x_i}}{x_1! \cdots x_n!} e^{-n\lambda},$$

und daraus sieht man, dass $x \mapsto \sum_i x_i$ eine suffiziente Statistik für dieses Modell ist.

5. Nun sei $\Omega = \mathbb{N}$, wir betrachten alle Gleichverteilungen $\mathbb{P}_a$ auf Mengen der Form $\{1, \ldots, a\}$. Unter $\mathbb{P}_a^n$ ist die Wahrscheinlichkeit für ein n-Tupel $(x_1, \ldots, x_n)$ gleich $1/a^n$, wenn alle x_i im Intervall liegen und Null sonst. Das kann man etwas gekünstelt als

$$h(x) g_a(\max x_i)$$

schreiben, dazu muss man $h = 1$ setzen und g_a als $1/a^n$ mal charakteristische Funktion von $\{1, \ldots, a\}$ definieren. Folgerung: $x \mapsto \max x_i$ ist suffizient für dieses Modell.

6. Nun kann man auch sagen, in welchem Sinne es „auf die Reihenfolge nicht ankommt“. Wir betrachten ein ganz beliebiges statistisches Modell. $T : \Omega^n \to \mathbb{R}^n$ soll diejenige Abbildung sein, die einfach der Größe nach sortiert. (Zur Erinnerung: Wir haben $\Omega \subset \mathbb{R}$ vorausgesetzt.) Zum Beispiel ist $T(3,2,2,4,1) = (1,2,2,3,4)$. Behauptung: T ist suffizient. Der Beweis ist leicht, denn wegen der Unabhängigkeit der Beobachtungen haben Permutationen die gleiche Wahrscheinlichkeit, man kann also einfach $h = 1$ setzen und g_θ als die Wahrscheinlichkeit von x unter $\mathbb{P}_\theta^n$ definieren.
Dieses T heißt übrigens die *Ordnungsstatistik*.

[8)] Dieses Beispiel wurde weiter oben schon einmal diskutiert.

7. In Analogie zu Beispiel 5 folgt: Betrachtet man auf $\mathbb{Z}$ alle Gleichverteilungen auf Intervallen der Form $\{a, a+1, \ldots, b\}$ mit $a < b$, so ist $x \mapsto (\min x_i, \max x_i)$ eine suffiziente Statistik.

Vollständigkeit

Wir benötigen noch eine etwas technische Bedingung. Gegeben ist eine suffiziente Statistik, und es soll ausgedrückt werden, dass die durch T induzierte Zerlegung von Ω^n in die Ω_y „nicht zu fein“ ist.

Definition 2.3.5. *Eine Statistik T heißt* vollständig*, wenn gilt: Ist $f : \Omega^n \to \mathbb{R}$ eine Funktion mit den Eigenschaften:*

(i) f ist auf jedem Ω_y konstant,

(ii) der Erwartungswert von f unter $\mathbb{P}_\theta^n$ ist Null für alle θ,

so ist f die Nullfunktion[9)].

Bemerkungen und Beispiele:

1. Mal angenommen, es gibt m verschiedene Ω_y. Bezeichnet man die f-Werte darauf als $f_1, \ldots, f_m$ und die Wahrscheinlichkeiten unter $\mathbb{P}_\theta$ mit $p_1^\theta, \ldots, p_m^\theta$, so heißt die Vollständigkeitsbedingung gerade: Aus

$$\sum_i f_i p_i^\theta = 0$$

(für alle θ) soll $f_1 = \cdots = f_m = 0$ folgen. Das bedeutet – das lernt man in der Linearen Algebra – dass die Vektoren

$$(p_1^\theta, \ldots, p_m^\theta),\ \theta \in \Theta,$$

ein Erzeugendensystem im $\mathbb{R}^m$ bilden müssen; Θ darf also nicht „zu klein“ sein.

2. Betrachte alle Bernoulliverteilungen $\mathbb{P}_p$ auf $\{0,1\}$, davon n-fache Kopien und die Statistik T = Summe; wir wissen schon, dass sie suffizient ist. Sei nun $f : \{0,1\}^n \to \mathbb{R}$ eine Funktion, die nur von $k = \sum x_i$ abhängt. Da es zu jedem möglichen k genau $\binom{n}{k}$ derartige x gibt, ist der Erwartungswert von f unter $\mathbb{P}_p^n$ gleich

$$\sum_k f(k) \binom{n}{k} p^k (1-p)^{n-k}.$$

Schreibt man das als

$$(1-p)^n \sum_k f(k) \binom{n}{k} \left(p/(1-p)\right)^k,$$

so sieht man: Ist dieser Ausdruck für alle p gleich Null, so muss wegen des Identitätssatzes für Polynome $f = 0$ gelten.

[9)] Genauer: $\mathbb{P}_\theta$-fast sicher gleich Null für alle θ.

3. Wir behaupten, dass die Ergebnissumme auch für das System aller Poissonverteilungen vollständig ist. Sei dazu f eine nur von dieser Summe abhängige Funktion. Der Erwartungswert unter $\mathbb{P}_\lambda^n$ ist

$$\sum_k f(k) \sum_{\sum x_i = k} \frac{\lambda^k}{x_1! \cdots x_n!} e^{-n\lambda}.$$

Wenn diese Summe für alle λ verschwindet, so muss f – wieder nach dem Identitätssatz für Potenzreihen – gleich Null sein.

4. Betrachte die Gleichverteilungen $\mathbb{P}_a$ auf allen $\{1, \ldots, a\}$. Behauptung: Das Maximum aus den n Stichprobenwerten ist vollständig. Sei f eine Funktion des Maximums, $f : \mathbb{N} \to \mathbb{R}$. Setzt man $m_k =$„Anzahl der Elemente des $\mathbb{N}^n$ mit Maximum k“, so ist zu zeigen, dass aus

$$\sum_{k=1}^{a} f(k) m_k / a^n = 0$$

(alle a) folgt, dass $f = 0$ ist. Das ist aber klar, man braucht nur sukzessive $a = 1, 2, \ldots$ einzusetzen.

5. Hat man für ein T die $\mathbb{P}_\theta^n$ gemäß Satz 2.3.4 geschrieben, so gilt: T ist genau dann vollständig, wenn gilt: Für jede auf dem Bild von T definierte Funktion F, für die $\sum F(y) g_\theta(y) = 0$ für alle θ gilt, ist $F = 0$.

Der Hauptsatz

Hier das Ergebnis unserer Bemühungen:

Satz 2.3.6. *(Lehmann-Scheffé) Gegeben seien ein statistisches Modell und eine Zielfunktion γ. T sei eine suffiziente und vollständige Statistik für dieses Modell.*

(i) Angenommen, es gibt einen erwartungstreuen Schätzer d (bei n Beobachtungen) für γ. Dann gibt es auch einen erwartungstreuen Schätzer d^* mit gleichmäßig bester Varianz, man kann ihn als $d^* = d_T \circ T$ wählen.

(ii) Umgekehrt: Ist es möglich, einen erwartungstreuen Schätzer der Form $d^* = d' \circ T$ zu finden, so ist das schon ein Schätzer mit gleichmäßig bester Varianz.

Beweis: Der Beweis ist erfreulich einfach.
(i) Sei $\tilde{d}$ ein weiterer erwartungstreuer Schätzer, wir können ihn verbessern, wenn wir zur Faktorisierung $\tilde{d}_T \circ T$ über T übergehen. Da $\tilde{d}_T \circ T$ und $d_T \circ T$ erwartungstreu sind, hat die Differenz unter allen $\mathbb{P}_\theta^n$ Erwartungswert 0, außerdem ist sie über T faktorisierbar. Damit muss – wegen der Vollständigkeit – $\tilde{d}_T \circ T = d_T \circ T$ gelten, insbesondere hat $d_T \circ T$ nicht schlechtere Varianzen als $\tilde{d}$.

(ii) Man zeigt einfach, dass $(d^*)_T = d'$ gilt: Das ist aufgrund der Definition von $(d^*)_T$ klar. Nach dem ersten Teil ist dann $d' \circ T$ bestmöglich. □

Bemerkungen und Beispiele:

1. Man kann sich die folgende **Faustregel** merken: Hat man eine vollständige und suffiziente Statistik für ein statistisches Modell gefunden, so ist alles ganz einfach. Man muss nur versuchen, einen erwartungstreuen Schätzer für γ so anzugeben, dass bei der Berechnung nicht x selbst, sondern nur Tx auftritt.

2. Die Ergebnisse zur Binomialverteilung und zur Poissonverteilung implizieren:

- Das unbekannte p bei einer Bernoulliverteilung schätzt man aus einer Stichprobe von n Versuchen am besten durch das Stichprobenmittel.
- Ist λ bei einer Poissonverteilung gesucht, so sollte man ebenfalls mit dem Stichprobenmittel schätzen.

3. Die vorstehenden Ergebnisse sind ziemlich plausibel, es gibt aber auch Schätzer, bei denen man ohne theoretischen Hintergrund keine Chance hat.

Wir betrachten noch einmal alle Gleichverteilungen auf allen diskreten Intervallen $\{1, \dots, a\}$ mit unbekanntem a. (Achtung: Das ist *nicht* das Quizbeispiel, da ging es um Intervalle $[\,0, a\,]$.) Es wird n mal gezogen, und man soll a schätzen.

Da wir schon wissen, dass $x \mapsto \max x_i$ eine suffiziente und vollständige Statistik ist, müssen wir nur einen erwartungstreuen Schätzer angeben, der eine Funktion von $y := \max x_i$ ist. Behauptung:

$$d(x_1, \dots, x_n) := \frac{y^{n+1} - (y-1)^{n+1}}{y^n - (y-1)^n}$$

ist so ein Schätzer. Zum Beweis fixieren wir ein a. Ist $1 \le b \le a$, so ist die Wahrscheinlichkeit, dass $\max x_i \le b$ gilt, offensichtlich gleich b^n/a^n. Folglich ist die Wahrscheinlichkeit, dass dieses Maximum exakt gleich b ist, gleich $\dfrac{b^n - (b-1)^n}{a^n}$, und für den Erwartungswert von d unter $\mathbb{P}_a^n$ ergibt sich

$$\begin{aligned} E_a(d) &= \sum_{y=1}^{a} \frac{y^{n+1} - (y-1)^{n+1}}{y^n - (y-1)^n} \cdot \frac{y^n - (y-1)^n}{a^n} \\ &= \sum_{y=1}^{a} \frac{y^{n+1} - (y-1)^{n+1}}{a^n} \\ &= a. \end{aligned}$$

Dieser Schätzer ist also bestmöglich.

> Um ihn besser zu verstehen, kann man den zweiten Mittelwertsatz der Differentialrechnung auf die Funktionen x^{n+1} und x^n auf dem Intervall $[\,y-1, y\,]$ anwenden. Damit folgt, dass d gleich $\dfrac{n+1}{n} z$ ist, wo z zwischen $y-1$ und y liegt.

Potenzreihenfamilien

Es ist recht lästig, immer wieder neu Suffizienz und Vollständigkeit nachzuprüfen. Deswegen ist es sinnvoll, ein für allemal eine große Klasse von statistischen Modellen einzuführen, für die man ein suffizientes T leicht finden kann.

Definition 2.3.7. *Es sei $\Omega \subset \mathbb{N}_0$. Weiter sei $\Theta \subset \mathbb{R}^+$ ein Intervall. Ein durch Maße $(\mathbb{P}_\theta)$ definiertes statistisches Modell heißt eine* Potenzreihenfamilie*, falls es Funktionen $t, h : \Omega \to \mathbb{R}$ und Zahlen c_θ so gibt, dass*

$$\mathbb{P}_\theta(\{x\}) = c_\theta \theta^{t(x)} h(x)$$

für alle x gilt.

Beispiele:

1. Die Bernoulliverteilungen sind ein erstes Beispiel, man muss allerdings etwas genauer hinsehen. Für $x = 0, 1$ ist doch

$$\mathbb{P}_\theta(\{x\}) = \frac{1}{1+\theta}\theta^x,$$

wenn man $\theta := p/(1-p)$ setzt. Geht man also zum Parameter θ über, so liegt eine Potenzreihenfamilie vor (mit $t(x) = x$, $h = 1$ und $c_\theta = 1/(1+\theta)$).

2. Die Familie der Poissonverteilungen auf $\mathbb{N}_0$ ist leicht als Potenzreihenfamilie zu erkennen: Es ist $P_\lambda(\{k\}) = \lambda^k e^{-\lambda}/k!$, und dieser Ausdruck ist gleich $c_\lambda \lambda^{t(k)} h(k)$, wenn wir definieren:

$$c_\lambda := e^{-\lambda},\ t(k) := k,\ h(k) := 1/k!\,.$$

Die Wichtigkeit der Definition resultiert aus dem

Satz 2.3.8. *Es werden Stichproben mit n Elementen gezogen. Definiert man dann T durch*

$$T(x_1, \ldots, x_n) := t(x_1) + \cdots + t(x_n),$$

so ist T eine suffiziente und vollständige Statistik.

Beweis: Die Suffizienz ist mit Satz 2.3.4 leicht einzusehen (beachte, dass die Stichproben $x_1, \ldots, x_n$ unabhängig gezogen werden):

$$\mathbb{P}^n_\theta(\{(x_1, \ldots, x_n)\}) \quad = \quad c^n_\theta \left(h(x_1) \cdots h(x_n)\right) \theta^{T(x)}.$$

Für die Vollständigkeit müssen wir von einer Funktion f ausgehen, die sich über T faktorisieren lässt, die also die Form $F \circ T$ hat. Es ist

$$\begin{aligned} E_\theta(f) &= \sum_x f(x)\mathbb{P}^n(\{x\}) \\ &= \sum_y \sum_{Tx=y} f(x)\mathbb{P}^n_\theta(\{x\}) \\ &= \sum_y F(y) \sum_{Tx=y} \mathbb{P}^n_\theta(\{x\}) \\ &= c^n_\theta \sum_y F(y)\theta^y \sum_{Tx=y} h(x_1) \cdots h(x_n). \end{aligned}$$

Die y sind dabei gewisse Elemente aus $\mathbb{R}$. Ist der Erwartungswert also für alle θ gleich Null, so folgt aus dem Identitätssatz für Potenzreihen, dass F und damit f verschwinden muss.

Nachtrag: Man kann die vorstehenden Überlegungen mit erheblichem Aufwand auf Wahrscheinlichkeitsräume übertragen, die durch Dichtefunktionen definiert sind. Hauptproblem dabei: Was sind denn bedingte Wahrscheinlichkeitsmaße?

> Intuitiv ist klar, was gemeint ist: Wenn z. B. (x, y) gleichverteilt auf dem Einheitskreis ist und $x = 0$ gilt, dann folgt daraus, dass y gleichverteilt in $[\,-1, 1\,]$ sein wird.

Wieder gibt es dann Statistiken, die, wenn sie suffizient und vollständig sind, sofort zu optimalen Schätzern führen.

Beispiele:

1. So ist $x \mapsto \max x_i$ eine suffiziente und vollständige Statistik für die Familie aller Gleichverteilungen auf allen $[\,0, a\,]$. Es folgt sofort, dass $x \mapsto \dfrac{n+1}{n} \max x_i$ der bestmögliche Schätzer für a ist. (D.h., dass wir das Quizproblem schon optimal gelöst hatten.)

2. Für die Gleichverteilungen auf allen Intervallen $[\,a, b\,]$ mit $a \leq b$ ist $x \mapsto (\max x_i, \min y_i)$ suffizient und vollständig. *So* gewinnt man leicht optimale Schätzer für a und b.

3. Wieder gibt es Potenzreihenfamilien, in die der richtige Schätzer schon eingebaut ist, alle gängigen Verteilungen gehören zu dieser Klasse. Es folgt zum Beispiel, dass das Stichprobenmittel die beste Möglichkeit liefert, den Erwartungswert von Normalverteilungen zu schätzen.

2.4 Ergänzungen

In diesem Abschnitt soll kurz auf weitere Themen eingegangen werden, die im Zusammenhang mit dem Schätzproblem interessant sind.

Bayes-Schätzer

Wir betrachten das übliche statistische Modell mit einer gegebenen Zielfunktion. Bisher hatten wir angenommen, dass die möglichen θ völlig gleichberechtigt auftreten. Das ist sicher zu einfach gedacht, wenn man schon einige Erfahrung mit diesem speziellen Testproblem hat: Man wird doch evtl. gewisse θ häufiger erwarten als andere. Mathematisch etwas formaler heißt das, dass wir ein Wahrscheinlichkeitsmaß α auf Θ postulieren, das unsere Erfahrung zusammenfasst; α heißt die *a-priori-Verteilung* des Problems. Und wieder gilt:

> *Informationen verändern Wahrscheinlichkeiten!*

Wenn wir nun ein- oder vielleicht sogar mehrfach abfragen, wird sich unsere Einschätzung möglicherweise ändern. Um das zu quantifizieren, braucht nur an den Satz von Bayes erinnert zu werden, *damit* sind die neuen Wahrscheinlichkeiten leicht angebbar.

Klar, dass man im Laufe der Zeit gelernt hat, das auf fast beliebigen Räumen zu beherrschen. Um aber hier die Ideen nicht zu sehr durch Technik zu überlagern, konzentrieren wir uns auf den Fall, in dem Θ und Ω endlich sind; das erspart uns wieder, bedingte Wahrscheinlichkeitsmaße zu definieren.

> Unsere Ausgangssituation ist die folgende:
>
> Gegeben sind $\Theta = \{1, \ldots, n\}$ und $\Omega = \{1, \ldots, m\}$ sowie Maße $\mathbb{P}_i$ auf Ω für $i \in \Theta$. Zusätzlich gibt es Zahlen $\alpha_i \geq 0$ mit $\sum \alpha_i = 1$, sie beschreiben das Wahrscheinlichkeitsmaß α. Mit
>
> $$p_{ij} := \mathbb{P}_i(\{j\})\alpha_i$$
>
> gilt dann: p_{ij} ist die Wahrscheinlichkeit, dass i ausgewählt und dann j gezogen wird.
>
> Umgekehrt kann man auch beliebige $p_{ij} \geq 0$ mit $\sum_i \sum_j p_{ij} = 1$ vorgeben und daraus die α_i und die $\mathbb{P}_i$ berechnen: $\alpha_i := \sum_j p_{ij}$, $\mathbb{P}_i(\{j\}) := p_{ij}/\alpha_i$.

Die $\mathbb{P}_i(\{j\})$ sind gerade die $\mathbb{P}(j|i)$; uns interessieren aber die $\mathbb{P}(i|j)$. Dabei steht $\mathbb{P}$ für dasjenige Maß, dass die Gesamtsituation – also „Auswählen von i gemäß $\alpha_1, \ldots, \alpha_n$, dann Auswählen von j gemäß $\mathbb{P}_i$“ – beschreibt; $\mathbb{P}$ ist ein Maß auf $\Theta \times \Omega$, und $\mathbb{P}(\{(i,j)\}) = p_{ij}$. Die Zahlen $\mathbb{P}(i|j)$ haben die folgende Interpretation: Wird konkret als Stichprobe j gezogen, so sollen wir Prognosen darüber anstellen, welches i dafür wohl verantwortlich war. Es ergibt sich so ein Maß auf Θ, das als *a-posteriori-Verteilung* unter der Bedingung j bezeichnet wird; wir schreiben dafür π^j, d.h., das Maß π^j ist durch die Zahlen $\pi^j_i := \mathbb{P}(i|j)$ gegeben. Nun wird die Bayes-Formel wichtig, danach ist

$$\pi^j_i = p_{ij}/\beta_j,$$

wobei $\beta_j := \sum_i p_{ij}$.

> *Beispiel:* Es sei $n = m = 2$ und $p_{11} = p_{22} = 1/8$ und $p_{12} = p_{21} = 3/8$. Damit ist $\alpha_1 = \alpha_2 = 1/2$.
>
> Nun wird $j = 2$ gemessen, $i = 1$ ist dadurch favorisiert. Und wirklich ergibt sich
>
> $$\pi^2_1 = 3/4,\ \pi^2_2 = 1/4.$$

Beim Schätzen soll unser Vorwissen natürlich auch eine Rolle spielen. Ist $\gamma : \Theta \to \mathbb{R}$ eine Zielfunktion (bei uns gegeben durch $\gamma_1, \ldots \gamma_n$), so soll γ durch einen Schätzer $d : \Omega \to \mathbb{R}$ (gegeben durch $d_1, \ldots, d_m$) geschätzt werden. Falls wirklich $\theta = i$ gilt, ist der quadratische Fehler durch $\sum_j (d_j - \gamma_i)^2 p_{ij}/\alpha_i$ gegeben.

Wichten wir ihn mit α_i und summieren auf, so ist das so etwas wie das erwartete Fehlerrisiko. Daher definieren wir $F_\alpha(d) := \sum_i \sum_j (d_j - \gamma_i)^2 p_{ij}$.

Man möchte natürlich so gut wie möglich schätzen: Ein Schätzer d heißt ein *Bayes-Schätzer*, wenn $F_\alpha(d)$ minimal ist. Überraschender Weise kann man sehr schnell Existenz und Eindeutigkeit zeigen, nebenbei fällt eine explizite Formel ab:

Satz 2.4.1. *Definiere einen Schätzer d^B punktweise als Erwartungswert über die a-posteriori-Verteilung, also durch die Formel*

$$d_j^B := E_{\pi^j}(\gamma) = \sum_i \gamma_i \pi_i^j = \sum_i \gamma_i p_{ij}/\beta_j.$$

Dann gilt: d^B ist der eindeutig bestimmte Bayes-Schätzer.

Bemerkung: d^B ist im Allgemeinen nicht erwartungstreu. Sind z.B. alle p_{ij} gleich, so ist d^B konstant.

Beweis: Im Beweis wird nur zu beachten sein, dass $\sum_i \gamma_i p_{ij} = \beta_j d_j^B$ und $\sum_i p_{ij} = \beta_j$ gilt. Sei d eine beliebige Schätzfunktion. Es ist

$$\begin{aligned} F_\alpha(d) - F_\alpha(d^B) &= \sum_i \sum_j \left((d_j - \gamma_i)^2 - (d_j^B - \gamma_i)^2 \right) p_{ij} \\ &= \sum_i \sum_j \left(d_j^2 - (d_j^B)^2 - 2\gamma_i (d_j - d_j^B) \right) p_{ij} \\ &= \sum_j \left(d_j^2 - (d_j^B)^2 - 2 d_j d_j^B + 2 (d_j^B)^2 \right) \beta_j \\ &= \sum_j \left(d_j - d_j^B \right)^2 \beta_j. \end{aligned}$$

Dieser Ausdruck ist ≥ 0, und er verschwindet genau dann, wenn $d = d^B$. Damit ist alles gezeigt.

> *Bemerkung:* Man kann auch mit analytischen Methoden zum gleichen Ergebnis kommen. Die Funktion F_α ist doch strikt konvex und nimmt ihr eindeutig bestimmtes Minimum an. Das findet man durch Nullsetzen des Gradienten, als Lösung des entsprechenden Gleichungssystems ergeben sich die Komponenten von d^B.
>
> Bei diesem Weg ist allerdings nicht offensichtlich, wie eine Verallgemeinerung auf den Fall beliebiger – nicht notwendig endlicher – Maßräume aussehen könnte.

□

Konfidenzbereiche

Wir wollen hier einen anderen Ansatz für das Schätzen besprechen. Bisher war es stets so, dass wir aufgrund der konkreten Stichprobe $x = (x_1, \ldots, x_n)$

ein $d(x) \in \mathbb{R}$ als konkreten Vorschlag für eine Schätzung von $\gamma(\theta)$ angegeben hatten: Man spricht dann von einer *Punktschätzung*. Das führt zum Beispiel beim Schätzen einer Wahrscheinlichkeit (fällt die Reißzwecke auf den Rücken?) zu Aussagen des Typs: „Die Schätzung ist 0.384“. Niemand glaubt dann, dass die gesuchte Wahrscheinlichkeit wirklich diesen Wert hat, doch wie kann man diese Ungewissheit angemessener beschreiben?

Intuitiv ist klar, dass die Sicherheit, mit der etwas behauptet wird, mit der Präzision einer Aussage abnehmen wird. In der Statistik verfährt man so, dass man als erstes eine Wahrscheinlichkeit festlegt, die man als Fehlerwahrscheinlichkeit zu tolerieren bereit ist. Sie heißt die *Irrtumswahrscheinlichkeit* und wird mit α bezeichnet. Typische Werte für α sind 0.1, 0.05 und 0.01, die Wahl wird von der speziellen Situation abhängen. Die Zahl $1 - \alpha$ ist dann die Wahrscheinlichkeit, mit der man auf eine richtige Antwort hoffen kann, sie heißt das *Konfidenzniveau*.

Angenommen, ein α ist fixiert. Wenn dann die Stichprobe den Wert x ergibt, möchte man eine „möglichst kleine “ Teilmenge Θ_x von $\mathbb{R}$ finden: *Das* ist unser Vorschlag für den Bereich, in dem $\gamma(\theta)$ zu finden ist. Wir wollen uns höchstens mit Wahrscheinlichkeit α irren, das führt zu der folgenden Forderung:

> Es sei θ der in Wirklichkeit ausgewählte Parameter. Dann werden die $x \in \Omega^n$ gemäß $\mathbb{P}_\theta^n$ erzeugt, jedes x liefert ein Θ_x, und mit Wahrscheinlichkeit $1 - \alpha$ soll $\gamma(\theta)$ darin enthalten sein. In Formeln: Für alle θ soll
>
> $$\mathbb{P}_\theta^n(\{x \mid \gamma(\theta) \notin \Theta_x\}) \leq \alpha$$
>
> sein[10)].

Wenn eine derartige Zuordnung gefunden wurde, spricht man von einer *Konfidenzbereichsschätzung mit Irrtumswahrscheinlichkeit* α (oder: zum Konfidenzniveau $1 - \alpha$).

Solche Verfahren haben wirklich die Eigenschaft, dass die Irrtumswahrscheinlichkeit kontrolliert werden kann, das ist in die Definition eingebaut. Formal sind auch sehr „feige“ Konfidenzbereichsschätzungen zugelassen: Man könnte zum Beispiel alle $\Theta_x = \mathbb{R}$ setzen, dann ist die Irrtumswahrscheinlichkeit sogar Null. Die Konfidenzmengen sollen aber zu vorgelegtem x möglichst klein sein, um möglichst präzise Aussagen zu erhalten. Der Einfachheit halber wählt man oft Intervalle (in $\mathbb{N}$ oder $\mathbb{R}$), man spricht dann auch von *Konfidenzintervallen*.

Es bleibt offen, wie man Konfidenzbereiche konkret findet. Wir diskutieren hier einige Beispiele, bei denen γ die identische Abbildung ist:

Beispiele:

1. Hier soll ein *allgemeines Verfahren* beschrieben werden. α sei vorgegeben. Wir wählen für jedes θ ein Ereignis $\Omega_{\theta,\alpha}$, so dass $\mathbb{P}_\theta(\Omega_{\theta,\alpha}) \geq 1 - \alpha$. (Im Idealfall sollte „=“ gelten, das ist im diskreten Fall aber nicht immer zu erreichen.)

[10)] Wir setzen wie immer stillschweigend voraus, dass die auftretenden Mengen messbar sind, wenn ein Maß darauf angewendet wird.

Nun können leicht Konfidenzbereiche gefunden werden: Definiere, für $x \in \Omega$, die Konfidenzmenge Θ_x durch

$$\Theta_x := \{\theta \mid x \in \Omega_{\theta,\alpha}\}.$$

Es ist dann durch die Definition sichergestellt, dass es sich wirklich um Konfidenzmengen handelt.

2. Als konkretes Beispiel dazu betrachten wir das *Binomialmodell.* Es ist also n fixiert, und bei unbekanntem p werden n Abfragen gemacht, die zu k Erfolgen führen. p soll zum Konfidenzniveau $1-\alpha$ geschätzt werden.

Man verfährt nach dem in „1.“ beschriebenen Verfahren. Fixiere p. Um ein möglichst kleines Konfidenzintervall zu erhalten, wählt man ein möglichst kleines Intervall in $\mathbb{N}$ der Form $\{a, a+1, \ldots, b\}$ mit Wahrscheinlichkeit $\geq 1-\alpha$. Das geht so:

> Suche k'_p möglichst groß mit
>
> $$\sum_{k=0}^{k'_p-1} b(k,n;p) < \alpha/2$$
>
> sowie k''_p möglichst klein mit
>
> $$\sum_{k''_p+1}^{n} b(k,n;p) < \alpha/2.$$
>
> Wenn man das für alle p berechnet hat und das Intervall von k'_p bis k''_p über p abträgt, ergibt sich eine Teilmenge von $[0,1] \times \{0,\ldots,n\}$, die von zwei aufwärts führenden „Treppen“ begrenzt ist.
>
> Nun muss man nur noch dieses Gebilde bei gegebenem k mit der zur x-Achse parallelen Geraden auf der Höhe k schneiden. Es ergibt sich ein Intervall $[\,p_u(k), p_o(k)\,]$, *das* ist das gesuchte Konfidenzintervall zu k.

Die Berechnung muss man übrigens nicht selbst durchführen, man kann die gesuchten Werte – also Intervalle des Typs $[\,p_u(k), p_o(k)\,]$ – aus Tafelwerken ablesen. Im Anhang ist auch eine entsprechende Tabelle abgedruckt.

Hier eine typische Anwendung, dabei seien $n = 50$ und $\beta = 0.99$ (also $\alpha = 1-\beta = 0.01$). Ergeben sich dann 11 Erfolge, so führt das zu dem Konfidenzintervall $[\,0.08, 0.42\,]$: Die hohe Sicherheit wird durch ein recht großes Konfidenzintervall erkauft.

3. Mal angenommen, wir haben eine Normalverteilung mit unbekanntem μ und bekanntem σ vor uns. Sie wird n-mal abgefragt, daraus soll ein Konfidenzintervall für μ gefunden werden. Wir argumentieren so:

- $x_1, \ldots . x_n$ sei das Ergebnis der Abfrage, mit $\overline{x}$ bezeichnen wir wie üblich das Stichprobenmittel. Es ist bekannt, dass $\overline{x}$ $N(\mu, \sigma^2/n)$ verteilt ist. Beachte: $(\overline{x} - \mu)\sqrt{n}/\sigma$ ist dann $N(0,1)$-verteilt.

- Wähle ein Intervall der Form $[\,-a, a\,]$, so dass

$$\mathbb{P}([\,-a, a\,]) = 1 - \alpha$$

  unter $N(0,1)$. (Das kann mit Hilfe einer Tafel von $N(0,1)$ leicht gefunden werden: Bestimme a so, dass $\mathbb{P}_{N(0,1)}(x \leq a) = 1 - \alpha/2$.)

- Mit Wahrscheinlichkeit $1-\alpha$ liegt $(\overline{x}-\mu)\sqrt{n}/\sigma$ in $[\,-a, +a\,]$. Das ist gleichwertig zu: Mit Wahrscheinlichkeit $1-\alpha$ liegt μ in $[\,\overline{x} - a\sigma/\sqrt{n}, \overline{x} + a\sigma/\sqrt{n}\,]$.

Und das heißt: $[\,\overline{x} - a\sigma/\sqrt{n}, \overline{x} + a\sigma/\sqrt{n}\,]$ ist ein Konfidenzintervall für μ zum Konfidenzniveau $1 - \alpha$.

Ein **Beispiel** dazu: Es sei $\alpha = 0.05$, $\sigma = 10$ und $n = 40$. Der Wert von a aus den vorstehenden Überlegungen ist dann 1.96. Damit ist $10 \cdot 1.96/\sqrt{40} = 3.1$ auszurechnen. Fazit: Das Konfidenzintervall für μ ist $[\,\overline{x} - 3.1, \overline{x} + 3.1\,]$.

Maximum-likelihood-Schätzer

Maximum-likelihood-Schätzer wurden im diskreten Fall schon in der elementaren Stochastik eingeführt. Die Idee war einfach:

> Gegeben sei ein statistisches Modell mit einem diskreten Ω. Es wird ein (uns unbekanntes) θ_0 ausgewählt, dann wird n-mal abgefragt. Das Ergebnis sei $(x_1, \ldots, x_n)$.
>
> Setze $p_\theta(x_1, \ldots, x_n) := \mathbb{P}_\theta^n(\{x_1, \ldots, x_n\})$. Man gibt dann als Schätzwert für θ_0 dasjenige θ an, für das $p_\theta(x_1, \ldots, x_n)$ maximal ist.

So schließt der gesunde Menschenverstand, und es ergeben sich Schätzungen, die sehr plausibel sind. (Als Beispiel wurde in der elementaren Vorlesung eine p-Schätzung bei einer Binomialverteilung behandelt.)

Das Verfahren ist auf den kontinuierlichen Fall nicht direkt übertragbar, da dort $\mathbb{P}_\theta^n(\{x_1, \ldots, x_n\}) = 0$ für alle θ sein wird. Man behilft sich dadurch, dass man sich an die naive Interpretation erinnert, nach der Ausgaben für diejenigen Werte am wahrscheinlichsten sein werden, für die die Dichtefunktion maximal ist. So gelangt man zu

Definition 2.4.2. *Gegeben sei ein statistisches Modell, dabei sei Ω ein Intervall, und die $\mathbb{P}_\theta$ seien durch eine Dichtefunktion f_θ gegeben. Zu schätzen sei θ, d.h. γ ist die Identität.*
Ein Schätzer $d : \Omega \to \Theta$ heißt ein Maximum-likelihood-Schätzer*, wenn gilt:*

$$f_{d(x)}(x) = \max f_\theta(x).$$

Bemerkungen und Beispiele:

1. Hat man als Stichprobe n unabhängige Abfragen zur Verfügung, so muss man sich daran erinnern, dass das Produktmaß auf Ω^n die Dichtefunktion $(x_1, \ldots, x_n) \mapsto f_\theta(x_1) \cdots f_\theta(x_n)$ hat. Es ist dann dasjenige θ zu finden, bei dem dieser Ausdruck maximal wird.

2. Maximum-likelihood-Schätzer müssen nicht existieren und im Fall der Existenz nicht eindeutig bestimmt sein.

3. Im differenzierbaren Fall findet man solche Schätzer mit elementarer Schulmathematik: Ableitung Null setzen usw.

4. Es sei $\Omega = \mathbb{R}^+$, wir betrachten alle Gleichverteilungen auf Intervallen des Typs $[\,0, a\,]$. Nun werde die Stichprobe $(x_1, \ldots, x_n)$ gezogen. Die zugehörige Dichtefunktion ist Null, falls $\max x_i > a$ und $1/a^n$ sonst. Folglich ist $\max x_i$ ein Maximum-likelihood-Schätzer.
Das zeigt, dass solche Schätzer nicht erwartungstreu sein müssen.

5. Im Binomialmodell ist „Erfolgsanzahl geteilt durch Versuchsanzahl“ ein Maximum-likelihood-Schätzer für p.

6. Wir betrachten nun alle Normalverteilungen $N(a, \sigma^2)$ und suchen einen Maximum-likelihood-Schätzer für (a, σ^2), falls die Stichprobe $(x_1, \ldots, x_n)$ vorliegt. Das läuft auf das folgende analytische Problem hinaus:

> Finde zu vorgegebenem Vektor $x := (x_1, \ldots, x_n) \in \mathbb{R}^n$ dasjenige Tupel (a, σ^2), für das
>
> $$f(a, v) := \frac{1}{(2\pi v)^{n/2}} \prod_i \exp[-(x_i - a)/2v] = \frac{1}{(2\pi v)^{n/2}} \exp\bigl(-\sum_i (x_i - a)^2/2v\bigr)$$
>
> maximal ist; dabei ist $v := \sigma^2$.

Sicher muss man dazu versuchen, $\sum (x_i - a)^2$ zu minimieren: Die Lösung kennen wir schon, es ist das Stichprobenmittel $\overline{x}$ (s. Satz 1.4.1). Und nun muss noch ein optimales v gefunden werden. Das findet man durch Nullsetzen der Ableitung des Logarithmus von f, also aus der Gleichung

$$0 = -\frac{n}{2v} + \frac{1}{v^2} \sum (x_i - \overline{x})^2.$$

Wir erhalten

$$v = \frac{1}{n} \sum_i (x_i - \overline{x})^2,$$

also bis auf den Faktor $n/(n-1)$ die Stichprobenvarianz[11)].

Fisher-Information

[11)] Auch dieser Schätzer ist also nicht erwartungstreu, vgl. Satz 2.2.3.

Wir haben schon Methoden kennen gelernt, um bestmögliche Schätzer zu finden. Wie gut ist aber „bestmöglich“? Man kann aus den Eigenschaften des statistischen Modells manchmal vorhersagen, dass die Varianz von Schätzfunktionen einen gewissen Wert nicht unterschreiten kann. Qualitativ kann man das Ergebnis plausibel finden, die technischen Einzelheiten (und warum man genau *diese* Definitionen wählt) sind aber nur schwer zu motivieren.

Wieder beschränken wir uns auf eine Situation, die einerseits reichhaltig genug ist, andererseits aber keine schwierigen Techniken erfordert. Hier ist unser Ausgangspunkt:

> Gegeben seien ein Intervall $\Omega = [\,a, b\,]$ und darauf definierte differenzierbare Funktionen f_θ für $\theta \in \Theta \subset \mathbb{R}$. Die f_θ sollen Wahrscheinlichkeitsdichten sein, das zugehörige Maß sei $\mathbb{P}_\theta$, *das* ist unser statistisches Modell.
>
> Als Zielfunktion ist $\gamma(\theta) := \theta$ gegeben, es soll also das Wahrscheinlichkeitsmaß identifiziert werden.

Vor der eigentlichen Definition gibt es einige *Vorüberlegungen*. Wenn irgendein θ fixiert ist und ein $x \in \Omega$ gemäß $\mathbb{P}_\theta$ als Stichprobe erzeugt wurde, soll θ durch eine maximum-likelihood-Schätzung gefunden werden. Es ist also unter allen θ dasjenige zu finden, für das $f_\theta(x)$ maximal ist.

Deswegen ist die partielle Ableitung von $f_\theta(x)$ nach θ von Interesse, die wird am Schätzwert verschwinden. Gleichwertig dazu ist das Verschwinden von $(\partial f_\theta(x)/\partial\theta)/f_\theta(x)$, das ist gerade die Ableitung nach θ von $\log f_\theta(x)$.

Das führt zur ersten Definition: Die *Likelihood-Funktion* wird durch $L(\theta, x) := \log f_\theta(x)$ erklärt, für die partielle Ableitung nach θ schreiben wir der Einfachheit halber $L'(\theta, x)$. Qualitativ sollte dann plausibel sein: Wenn x fixiert ist, so suchen wir doch ein $\tilde\theta$ mit $L'(\tilde\theta, x) = 0$; und wenn $L'(\theta, x)$ „in der Regel groß“ ist, wird das zu „scharfen“ Schätzungen für θ führen. Das ist, zugegeben, etwas vage. Mindestens sollte es motivieren, dass die Größe

$$I(\theta) := E_\theta([L'(\theta, x)]^2)$$

wichtig sein könnte und dass ein großes $I(\theta)$ auf eine gute Schätzbarkeit hoffen lässt. Die Funktion $\theta \mapsto I(\theta)$ heißt die *Fisher-Information*, mit ihr wird wirklich eine quantitative Abschätzung gelingen.

Vorbereitungen: Sei d ein erwartungstreuer Schätzer für γ.

1. Die Erwartungstreue heißt, dass $\int_a^b d(x) f_\theta(x) dx = \theta$ für alle θ. Leitet man das partiell nach θ ab, so folgt

$$\begin{aligned} 1 &= \int_a^b d(x) \frac{\partial f_\theta(x)}{\partial\theta} dx \\ &= \int_a^b d(x) L'(\theta, x) f_\theta(x) dx \\ &= E_\theta(d(x) L'(\theta, x)). \end{aligned}$$

2. Für jedes θ ist $\int f_\theta(x)dx = 1$, da die f_θ Dichtefunktionen sind. Leitet man das partiell nach θ ab, folgt $\int f'_\theta(x)dx = 0$; wieder bezeichnet der Ableitungsstrich die partielle Ableitung nach θ.

3. Damit kann man schließen:

$$\begin{aligned} 0 &= \int f'_\theta(x)dx \\ &= \int L'_\theta(x) f_\theta(x)dx \\ &= E_\theta(L'). \end{aligned}$$

Und deswegen ist

$$\begin{aligned} Var_\theta(L') &= E_\theta([L']^2) - (E_\theta[L'])^2 \\ &= I(\theta). \end{aligned}$$

4. Nun betrachten wir die (bzgl. $\mathbb{P}_\theta$ quadrat-integrablen) Funktionen auf Ω als Vektorraum und definieren darauf ein Skalarprodukt durch

$$\langle f, g\rangle := \int f g d\mathbb{P}_\theta.$$

Die bisherigen Ergebnisse besagen dann:

- $\langle L', c\rangle = 0$ für jede konstante Funktion c.
- $\langle L', L'\rangle = I(\theta)$.
- $\langle d, L'\rangle = 1$.

5. Wir wenden noch für *dieses* Skalarprodukt die Cauchy-Schwarzsche Ungleichung an:

$$\begin{aligned} 1 &= \langle d, L'\rangle^2 \\ &= \langle d - \theta, L'\rangle^2 \\ &\leq ||d - \theta||^2 ||L'||^2 \\ &= Var_\theta(d) I_\theta. \end{aligned}$$

Damit haben wir die so genannte *Rao-Cramer-Informationsungleichung* bewiesen:

Satz 2.4.3. *Es gilt $Var_\theta(d) \geq 1/I_\theta$ für alle erwartungstreuen Schätzer d. Damit wird eine theoretische untere Schranke für die mögliche Güte von Schätzern gegeben. Besser als der Kehrwert der Fisher-Information kann die Varianz eines Schätzers niemals werden.*

Stochastische Schätzer

Für uns war ein Schätzer eine Funktion d, die einer Stichprobe $(x_1, \ldots, x_n) \in \Omega^n$ eine reelle Zahl zuordnet: *Das* ist unser Vorschlag für den zu schätzenden Aspekt $\gamma(\theta)$. Manchmal ist dieser Ansatz zu eng. Um die Verallgemeinerung, die wir gleich kennen lernen werden, zu motivieren, betrachten wir ein Beispiel:

> Das statistische Modell bestehe aus allen Gleichverteilungen auf Mengen der Form $\{b, b+1\}$ mit $b \in \mathbb{N}$, das zugehörige Maß soll mit $\mathbb{P}_b$ bezeichnet werden. Aus n Beobachtungen $x_1, \ldots, x_n$ soll b geschätzt werden.
>
> Dann ist klar: Kommen in der Stichprobe zwei verschiedene Zahlen vor, so wird die kleinere das gesuchte b sein. Wenn aber $x_1 = \cdots = x_n =: c$, so kann $b = c$ oder $b = c-1$ gelten, und beide Möglichkeiten sind völlig gleichberechtigt. Da könnte man auf die Idee kommen, sich durch Werfen einer Münze für eine von ihnen zu entscheiden.

Möchte man diese Idee formalisieren, so gelangt man zur

Definition 2.4.4. *Ein* stochastischer Schätzer für den Aspekt γ *ist eine Abbildung d^s, die jedem $x = (x_1, \ldots, x_n) \in \Omega^n$ ein Wahrscheinlichkeitsmaß Q_x auf $\mathbb{R}$ zuordnet.*

Die Interpretation: Liegt die Stichprobe x vor, so erzeuge eine reelle Zahl gemäß Q_x. Das *soll dann der Vorschlag für $\gamma(\theta)$ sein.*

Bemerkungen und Beispiele:

1. Die üblichen Schätzer d tauchen als Spezialfall auf: Da ist Q_x das Punktmaß bei $d(x)$.

2. Genau genommen muss man noch einige technische Feinheiten berücksichtigen, um vernünftig mit solchen stochastischen Schätzern arbeiten zu können. So muss man, um für d^s wieder einen Erwartungswert definieren zu können, gewisse Messbarkeitsvoraussetzungen machen. (Das ist allerdings nur im nicht-diskreten Fall von Bedeutung.)

3. Im vorstehenden Motivationsbeispiel hatten wir Q_x so definiert: Enthält x zwei verschiedene Zahlen als Komponenten, so ist Q_x das Punktmaß auf der kleineren, andernfalls – falls $x_i = c$ für alle i – ist Q_x die Gleichverteilung auf $\{c-1, c\}$.

Das Thema wird im nächsten Kapitel unter dem Namen „stochastische Tests“ noch einmal aufgegriffen werden. Da wird dann auch klar werden, dass optimale Lösungen in gewissen Fällen nur mit der Hilfe des Zufalls gefunden werden können.

2.5 Der Spezialfall normalverteilter Messungen

Im Fall normalverteilter Messungen lassen sich sehr präzise Aussagen machen. Das liegt daran, dass man die auftretenden Verteilungen in diesem Spezialfall

explizit berechnen kann. Die Ergebnisse werden auch in Kapitel 4 eine wichtige Rolle spielen.
Der Aufbau dieses Abschnitts ist wie folgt:

- Zunächst gibt es einen Nachtrag zur elementaren Stochastik. Wir führen zwei Verteilungsfamilien – die β- und die Γ-Verteilungen – ein.
- Dann behandeln wir einige Vorbereitungen: Wie verändern sich Dichtefunktionen, wenn man zu Funktionen der Zufallsvariablen übergeht.
- Anschließend studieren wir, wie sich einige natürliche Verteilungen ergeben, wenn man mit Normalverteilungen startet. Diese Ergebnisse sind fundamental für die Statistik.
- Diese Ergebnisse können nun leicht angewendet werden. Es wird kein Problem sein, alle möglichen Konfidenzintervalle bei unterschiedlichster Vorinformation zu berechnen.

Zwei weitere Verteilungsfamilien

In der Analysis ist in der Abteilung „Anwendungen der Integralrechnung“ auch die Γ-*Funktion* behandelt worden. Sie ist durch

$$\Gamma(r) := \int_0^\infty t^{r-1}e^{-t}dt$$

für $r > 0$ definiert. Man konnte sich leicht davon überzeugen, dass das Integral wirklich existiert und dass gilt:

- $\Gamma(1) = 1$.
- $\Gamma(r+1) = r\Gamma(r)$.
- $\Gamma(r) = (r-1)!$ für $r \in \mathbb{N}$.

Substituiert man (bei gegebenem $\alpha > 0$) $\alpha x = t$, so folgt

$$\Gamma(r) = \int_0^\infty \alpha^r x^{r-1}e^{-\alpha x}dx.$$

Und das heißt: Definiert man eine Funktion $\gamma_{\alpha,r}$ auf $[\,0,+\infty\,[$ durch

$$\gamma_{\alpha,r}(x) := \frac{\alpha^r}{\Gamma(r)}x^{r-1}e^{-\alpha x},$$

so ist das eine Dichtefunktion. Das dadurch definierte Wahrscheinlichkeitsmaß auf $[\,0,+\infty\,[$ heißt die *Gammaverteilung* $\Gamma_{\alpha,r}$.
(Implizit tauchte sie in der elementaren Wahrscheinlichkeitsrechnung schon einmal auf, als es um unabhängige Summen von Exponentialverteilungen ging. Auch ist $\Gamma_{\lambda,1}$ die Exponentialverteilung zum Parameter λ.)

Nun seien zwei Zahlen $a, b > 1$ vorgegeben. Dann definiert man

$$B(a,b) := \int_0^1 s^{a-1}(1-s)^{b-1}\,ds,$$

B heißt die *Eulersche Beta-Funktion*. Mit Hilfe von B kann man eine Dichtefunktion auf $[\,0,1\,]$ definieren. Man setzt

$$\beta_{a,b}(s) := \frac{s^{a-1}(1-s)^{b-1}}{B(a,b)},$$

der zugehörige Wahrscheinlichkeitsraum heißt die *Beta-Verteilung* $\beta_{a,b}$ auf $[\,0,1\,]$.

Technische Vorbereitungen

Hier sollen einige Techniken zusammengestellt werden, die im Folgenden benötigt werden. Einige sind schon in der elementaren Stochastik behandelt worden.

- Sind X und Y unabhängige Zufallsvariable mit Werten in $\mathbb{R}^+$, die eine Dichte f bzw. g haben, so hat auch $X+Y$ eine Dichte h. Sie ist gerade die Faltung von f und g, also

$$h(x) = (f * g)(x) = \int_0^x f(t)g(x-t)dt.$$

- X, Y seien reellwertige Zufallsvariable mit Dichten f, g. Sie sind genau dann unabhängig, wenn die vektorwertige Zufallsvariable (X,Y) die Dichtefunktion $(x,y) \mapsto f(x)g(y)$ hat, also genau dann, wenn die gemeinsame Dichte Produktform hat.

- Seien $B, C \subset \mathbb{R}^d$ offene Teilmengen. Auf B sei eine Dichtefunktion f definiert, wir fassen den zugehörigen Wahrscheinlichkeitsraum als induzierte Dichte der vektorwertigen Zufallsvariable $X =$ Identität auf.

  Weiter sei $\varphi : B \to C$ bijektiv, und φ und φ^{-1} sollen differenzierbar sein. Wir betrachten die Zufallsvariable $Y := \varphi(X)$.

  > Y entsteht also dadurch, dass zunächst ein Punkt x mit der f-Dichte erzeugt wird; ausgegeben wird dann $Y(x)$. Die Bildwerte liegen offensichtlich in C.

  Behauptung: Das Bildmaß von Y hat eine Dichte g, sie ist durch

$$g(y) := f(\varphi^{-1}(y))|\det J_{\varphi^{-1}}(y)|$$

  gegeben[12)].

[12)] J_f bezeichnet die Jacobimatrix einer Funktion f.

Beweis: Sei $D \subset C$. Schreibt man $D = \varphi(E)$, so wird genau dann Y in D liegen, wenn X in E liegt. Dafür kennen wir die Wahrscheinlichkeit, sie ist gleich $\int_E f$. Nach dem Transformationssatz für Integrale kann man das als

$$\int_D g(y)dy$$

schreiben, und damit ist die Behauptung schon bewiesen.

Im Spezialfall $d = 1$ wurde die Aussage schon in der elementaren Stochastik gezeigt: Ist f eine Dichte auf $[\,a, b\,]$ und ist $\varphi : [\,a, b\,] \to [\,c, d\,]$ streng monoton steigend, bijektiv und differenzierbar, so hat das durch φ auf $[\,c, d\,]$ induzierte Wahrscheinlichkeitsmaß die Dichte $f(\varphi^{-1})(\varphi^{-1})'(y)$; das folgt sofort aus der Substitutionsregel.

Verteilungen, die sich aus der Normalverteilung ergeben

Hier soll gezeigt werden, wie die eben eingeführten Verteilungen mit der Normalverteilung zusammenhängen. Das wird für das Folgende grundlegend sein.

Lemma 2.5.1. *Sei X $N(0, 1)$-verteilt, das induzierte Maß hat also die Dichte der Standard-Normalverteilung. Dann ist X^2 wie $\Gamma_{1/2,1/2}$ verteilt.*

Beweis: Wir betrachten zunächst $|X|$. Das ist eine $\mathbb{R}^+$-wertige Zufallsvariable, die dort aus Symmetriegründen die doppelte Dichte der Standardnormalverteilung hat, also

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}.$$

Wir interessieren uns doch für $\varphi(|X|)$, wo $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ durch $\varphi(x) := x^2$ definiert ist. Damit ist $\varphi^{-1}(y) = \sqrt{y}$, *das* muss in das Zweifache der Dichte der Standardnormalverteilung eingesetzt werden, anschließend ist mit der Ableitung zu multiplizieren. Als Dichtefunktion erhalten wir

$$\frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{\sqrt{y}}.$$

Das ist bis auf den Faktor $\Gamma(1/2)/\sqrt{\pi}$ die Dichte $\gamma_{1/2,1/2}$. Nun sind aber zwei Dichten, die bis auf einen Faktor übereinstimmen, identisch, außerdem muss der Faktor Eins sein. Es folgt die Behauptung, und als Nebenergebnis haben wir noch die bemerkenswerte Gleichung

$$\Gamma(1/2) = \sqrt{\pi}$$

erhalten. □

Satz 2.5.2. *Es seien $\alpha, r, s > 0$, und X bzw. Y seien unabhängige Zufallsvariable, die $\Gamma_{\alpha,r}$- bzw. $\Gamma_{\alpha,s}$-verteilt sind.*

(i) $X+Y$ *und* $X/(X+Y)$ *sind unabhängig.*

(ii) $X+Y$ *ist* $\Gamma_{\alpha,r+s}$*-verteilt.*

(iii) $X/(X+Y)$ *ist* $\beta_{r,s}$*-verteilt; insbesondere ist diese Verteilung unabhängig von* α*.*

Beweis: Die gemeinsame Verteilung von (X,Y) hat wegen der Unabhängigkeit die auf $]\,0,+\infty\,[^2$ definierte Dichte

$$(x,y)\mapsto \rho(x,y) := \gamma_{\alpha,r}(x)\gamma_{\alpha,s}(y) = \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)}x^{r-1}y^{s-1}e^{-\alpha(x+y)}.$$

Definiere

$$\varphi : B :=]\,0,+\infty\,[^2 \to]\,0,+\infty\,[\times]\,0,1\,[=: C$$

durch $\varphi(x,y) := (x+y, x/(x+y))$. Damit ist $\varphi(X,Y) = (X+Y, X/(X+Y))$, und wir können mit Vorbereitung 3 die gemeinsame Dichte von $(X+Y, X/(X+Y))$ ausrechnen.

Dazu stellen wir fest, dass φ^{-1} die explizite Form $(u,v)\mapsto(uv, u(1-v))$ hat. Das führt zur Jacobimatrix

$$J_{\varphi^{-1}}(u,v) = \begin{pmatrix} v & u \\ 1-v & -u \end{pmatrix}$$

mit der Determinante u. Das bedeutet für die Dichte von $(X+Y, X/(X+Y))$ die Formel

$$\begin{aligned}\rho(uv,u(1-v)) &= \frac{\alpha^{r+s}}{\Gamma(r)\Gamma(s)}u^{r+s-1}e^{-\alpha u}v^{r-1}(1-v)^{s-1}\\ &= \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}B(r,s)\gamma_{\alpha,r+s}(u)\beta_{r,s}(v).\end{aligned}$$

Das hat mehrere Konsequenzen. Erstens hat die Dichte Produktform, d.h. $X+Y$ und $X/(X+Y)$ sind wirklich unabhängig. Zweitens muss der Vorfaktor Eins sein (da ja γ und β Dichtefunktionen sind), wir erhalten also als Nebenresultat die Beziehung

$$B(r,s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}.$$

Und drittens kann man ablesen, dass $X+Y$ und $X/(X+Y)$ die gewünschten Dichten haben. □

Korollar 2.5.3. *Es ist* $\gamma_{\alpha,r} * \gamma_{\alpha,s} = \gamma_{\alpha,r+s}$*.*

Für uns am wichtigsten ist die nachstehende Folgerung:

Korollar 2.5.4. *Es seien* $X_1,\ldots,X_n$ *unabhängige* $N(0,1)$*-verteilte Zufallsvariable. Dann ist* $X_1^2+\cdots+X_n^2$ *wie* $\Gamma_{1/2,n/2}$ *verteilt.*

Für diese Verteilung gibt es einen eigenen Namen:

Definition 2.5.5. *Unter der* Chiquadrat-Verteilung mit n Freiheitsgraden *(kurz χ^2_n-Verteilung) versteht man die $\Gamma_{1/2,n/2}$-Verteilung. Sie ist auf $]\,0,+\infty\,[$ definiert und hat die Dichte*

$$\chi^2_n(x) = \gamma_{1/2,n/2}(x) = \frac{x^{n/2-1}}{\Gamma(n/2)2^{n/2}} e^{-x/2}.$$

Wir benötigen noch weitere Dichten:

Satz 2.5.6. *Es seien $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ unabhänige $N(0,1)$-verteilte Zufallsvariable. Betrachte*

$$W := \frac{1}{m}\sum X_i^2 \Big/ \frac{1}{n}\sum Y_j^2.$$

W hat auf $]\,0,+\infty\,[$ die Dichtefunktion

$$f_{m,n}(x) = \frac{m^{m/2}n^{n/2}}{B(m/2,n/2)} \frac{x^{m/2-1}}{(n+mx)^{(n+m)/2}}.$$

Beweis: Setze $X := \sum X_i^2$ und $Y := \sum Y_j^2$. Diese Zufallsvariablen sind unabhängig, und wir kennen die Dichten, nämlich $\gamma_{1/2,m/2}$ und $\gamma_{1/2,n/2}$. Nach Satz 2.5.2 ist dann $Z := X/(X+Y)$ Beta-verteilt, die Dichte ist $\beta_{m/2,n/2}$.

Wir schreiben nun W als

$$W = \frac{n}{m}\frac{X}{Y} = \frac{n}{m}\frac{Z}{1-Z} =: \varphi(Z).$$

Da $\varphi :]\,0,1\,[\to]\,0,+\infty\,[,\ x \mapsto (n/m)(x/(1-x))$ eine differenzierbare Bijektion ist, können wir die Dichte von W mit den Vorbereitungen explizit bestimmen. (Hier ist $d = 1$, die Jacobideterminante ist dann einfach die Ableitung.) Es ist $\varphi^{-1}(y) = my/(n+my)$ und damit

$$\big(\varphi^{-1}\big)'(y) = \frac{nm}{(n+my)^2}.$$

So ergibt sich als Dichte für W die Funktion

$$\begin{aligned}\beta_{m/2,n/2}\big(\varphi^{-1}(y)\big)\big(\varphi^{-1}\big)'(y) \quad &= \\ &= \frac{1}{B(m/2,n/2)}\Big(\frac{my}{n+my}\Big)^{m/2-1}\Big(\frac{n}{n+my}\Big)^{n/2-1}\frac{nm}{(n+my)^2} \\ &= \frac{1}{B(m/2,n/2)} \cdot \frac{m^{m/2}n^{n/2}y^{m/2-1}}{(n+my)^{n/2+m/2}}.\end{aligned}$$

Das ist gerade die Funktion $f_{m,n}$, und damit ist der Satz bewiesen. □

Definition 2.5.7. *Die eben durch die Dichtefunktion $f_{m,n}$ beschriebene Verteilung auf $]\,0,+\infty\,[$ heißt die* F-Verteilung mit m und n Freiheitsgraden*, sie wird kurz mit $F_{m,n}$ bezeichnet*[13)].

Die letzte für uns wichtige Verteilung wird im folgenden Satz eingeführt:

Satz 2.5.8. *Es seien $X, Y_1, \ldots, Y_n$ unabhängige $N(0,1)$-verteilte Zufallsvariable. Betrachte $W := X/\sqrt{(1/n)\sum Y_i^2}$.*
Dann hat W eine Dichtefunktion, nämlich die auf $\mathbb{R}$ definierte Funktion

$$\tau_n(x) := \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} \Big/\, B(1/2, n/2)\sqrt{n}.$$

Beweis: W^2 ist nach dem vorigen Satz $F_{1,n}$-verteilt. Damit ist die Dichte von $|W| = \sqrt{W^2}$ mit den üblichen Methoden leicht zu berechnen, es ergibt sich die Funktion $2y f_{1,n}(y^2)$. Beachte noch, dass W symmetrisch ist. Die Dichte von $|W|$ ist also nur zu spiegeln und durch zwei zu teilen. Es folgt: W hat die Dichte $f_{1,n}(y^2)|y| = \tau_n(y)$. □

Definition 2.5.9. *Die eben eingeführte Verteilung heißt die* Studentsche t-Verteilung mit n Freiheitsgraden*. Man spricht von der t_n-Verteilung.*

Die eben eingeführten Verteilungen sind deswegen wichtig, weil sie sich in natürlicher Weise bei der Untersuchung von Normalverteilungen ergeben. Eine ganz besondere Rolle wird der folgende Satz spielen:

Satz 2.5.10. *Es seien $a \in \mathbb{R}$ und $\sigma^2 > 0$. Gegeben seien $N(a, \sigma^2)$-verteilte unabhängige Zufallsvariable $X_1, \ldots, X_n$. Definiere dann*

$$M := \frac{1}{n}\sum X_i,\ V^* := \frac{1}{n-1}\sum (X_i - M)^2;$$

das sind gerade diejenigen Zufallsvariablen, die dem Stichprobenmittel und der Stichprobenvarianz entsprechen. Dann gilt:

(i) M und V^* sind unabhängig.

(ii) M ist $N(a, \sigma^2/n)$-verteilt, und $\dfrac{n-1}{\sigma^2}V^*$ ist χ^2_{n-1}-verteilt.

(iii) $\dfrac{\sqrt{n}(M-a)}{\sqrt{V^*}}$ ist t_{n-1}-verteilt.

Bemerkung: Der Satz ist überraschend: Erstens hätte man die Unabhängigkeit in (i) nicht erwartet, und zweitens ist nicht selbstverständlich, dass man konkrete Formeln für die Dichten angeben kann.

Beweis: Ist X_i $N(a, \sigma^2)$-verteilt, so ist $Y_i := (X_i - a)/\sigma$ $N(0,1)$-verteilt. Bezeichne die zu $Y_1, \ldots, Y_n$ konstruierten Zufallsvariablen analog mit M_Y und V_Y^*,

[13)] Sie ist nach dem Statistiker R.A. Fisher benannt.

wir nehmen an, dass die Aussage für den Fall $N(0,1)$-wertiger Variable schon bewiesen ist.

Dann sind also M_Y und V_Y^* unabhängig, das gilt dann auch für $M = \sigma M_Y + a$ und $V^* = \sigma^2 V_Y^*$. Weiter: Wenn $(n-1)V_Y^*$ die Verteilung χ^2_{n-1} hat, so auch $\frac{n-1}{\sigma^2}V^*$, denn diese beiden Zufallsvariablen stimmen überein. Und da

$$\frac{\sqrt{n}M_Y}{\sqrt{V_Y^*}} = \frac{\sqrt{n}(M-a)}{\sqrt{V^*}},$$

folgt aus der Gültigkeit von (iii) für die Y_i die entsprechende Aussage für die X_i.

Fazit: Wir dürfen o.B.d.A. annehmen, dass $a = 0$ und $\sigma = 1$ gilt.

Bevor es weitergeht, erinnern wir uns an die Definition einer *orthogonalen Matrix.* Das ist eine Matrix O mit der Eigenschaft $OO^\top = Id$. Orthogonale Matrizen entsprechen den linearen Abbildungen auf dem $\mathbb{R}^n$, die die Länge von Vektoren erhalten: Für jedes x gilt also $||x|| = ||Ox||$.

Seien nun $Z_1, \ldots, Z_n$ unabhängige $N(0,1)$-verteilte Zufallsvariable. Die gemeinsame Verteilung auf dem $\mathbb{R}^n$ ist

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-(x_1^2+\cdots+x_n^2)/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-||x||^2/2}.$$

Nach der Vorbemerkung folgt: Ist O eine orthogonale Matrix und definiert man $W_1, \ldots, W_n$ durch

$$(W_1, \ldots, W_n)^\perp := O(Z_1, \ldots, Z_n)^\perp,$$

so hat $(W_1, \ldots, W_n)$ die gemeinsame Dichtefunktion

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-||O^\top x||^2/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-||x||^2/2}.$$

(Hier war noch wichtig, dass die Jacobideterminante der Abbildung $x \mapsto O^\top x$ gleich Eins ist.)
Wichtige Folgerung: Auch die $W_1, \ldots, W_n$ sind unabhängige $N(0,1)$-verteilte Zufallsvariable.

Zurück zum Hauptbeweis. Die X_i sind unabhängig und (o.B.d.A.) $N(0,1)$-verteilt. Wir wählen irgendeine orthogonale Matrix O, für die die erste Zeile gleich

$$(1/\sqrt{n}, \ldots, 1/\sqrt{n})$$

ist[14)].

[14)] So etwas gibt es, man muss nur diesen Vektor zu einem Orthonormalsystem erweitern.

Definiere $Y_1, \ldots, Y_n$ durch $(Y_1, \ldots, Y_n)^\perp := O(X_1, \ldots, X_n)^\perp$. Nach Vorbemerkung sind die Y_i unabhängig und $N(0,1)$-verteilt, außerdem gilt nach Konstruktion

$$Y_1 = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) = \sqrt{n}M.$$

Wir beachten noch, dass $||X||^2 = ||Y||^2$, da O orthogonal ist[15]. Es folgt

$$\begin{aligned}(n-1)V^* &= \sum_{i=1}^{n}(X_i - M)^2 \\ &= \sum_{i=1}^{n} X_i^2 - nM^2 \\ &= ||X||^2 - nM^2 \\ &= ||Y||^2 - nM^2 \\ &= ||Y||^2 - Y_1^2 \\ &= \sum_{i=2}^{n} Y_i^2.\end{aligned}$$

Damit sind wir überraschenderweise schon fertig:

- $M = Y_1/\sqrt{n}$ und $V^* = (\sum_{i=2}^{n} Y_i^2)/(n-1)$ sind unabhängig.
- Dass M die behauptete Verteilung hat, ist klar, für die Verteilungsbehauptung von V^* muss man Korollar 2.5.4 zitieren.
- Dass $\sqrt{n}M/\sqrt{V^*}$ wie behauptet verteilt ist, folgt nun aus Satz 2.5.8.

Damit ist der Satz vollständig bewiesen. □

Anwendung: Konfidenzintervalle beim Gaußmodell

Die vorstehend bewiesenen Resultate sollen nun angewendet werden, um Konfidenzintervalle zu unterschiedlichen Schätzproblemen zu berechnen. Die für die konkrete Rechnung notwendigen Wahrscheinlichkeiten müssen aus Tafelwerken (oder einem Statistik-Hilfsprogramm) entnommen werden, da die auftretenden Integrale nicht geschlossen ausgewertet werden können. Im *Anhang* sind die wichtigsten *Tafeln* zu finden.

A. *Mittelwert einer Normalverteilung bei bekannter Streuung*

Problem: Das statistische Modell bestehe aus allen $N(a, \sigma^2)$, wobei σ bekannt ist. Das Problem kennen wir bereits: es wurde schon in Abschnitt 2.4 auf Seite 44 beschrieben, wie ein Konfidenzintervall zu finden ist.

B. *Mittelwert einer Normalverteilung bei unbekannter Streuung*

[15] X und Y sind die aus den X_i und den Y_i zusammengesetzten vektorwertigen Zufallsvariablen.

Problem: Diesmal besteht das Modell aus allen $N(a, \sigma^2)$, wobei σ fest, aber unbekannt ist. Gesucht ist ein Konfidenzintervall für a aus einer Stichprobe des Umfangs n. (Beispiel: Eine wirkliche Länge soll aus einer mehrfachen Längenmessung mit einem unbekannten Messgerät ermittelt werden).

Lösung: Bezeichne wie üblich mit α die Irrtumswahrscheinlichkeit. Bestimme dann mit Hilfe einer Tabelle ein r so, dass das Intervall $[\,-r, r\,]$ unter t_{n-1} die Wahrscheinlichkeit $1-\alpha$ hat. Löse dann noch

$$\frac{\sqrt{n}(M-a)}{\sqrt{V^*}} \in [\,-r, r\,]$$

nach a auf (wobei für M bzw. V^* das Stichprobenmittel bzw. die Stichprobenvarianz einzusetzen ist):

$$a \in \left[\, M - r \cdot \frac{\sqrt{V^*}}{\sqrt{n}}, M + r \cdot \frac{\sqrt{V^*}}{\sqrt{n}} \,\right].$$

Das rechts stehende Intervall ist das gesuchte Konfidenzintervall.

Begründung: Die „wirkliche“ Zufallsvariable ist $N(a, \sigma^2)$-verteilt, nach Satz 2.5.10 ist $\dfrac{\sqrt{n}(M-a)}{\sqrt{V^*}}$ t_{n-1}-verteilt.

Beispiel: Es sei $\alpha = 0.05$, $n = 10$, $\overline{x} = 3.1$ und $V_x = 12.3$. Wir müssen in der Tabelle mit 9 (!) Freiheitsgraden in der Spalte $t_{0.025}$ nachsehen, denn es sind die $\mathbb{P}(\{x \leq t\})$ tabelliert. Wir erhalten den Wert 2.262, d.h.: Unter t_9 hat $[\,-2.262, 2.262\,]$ die Wahrscheinlichkeit 0.95. Wir berechnen noch $2.262\sqrt{12.3}/\sqrt{10} = 2.50$. Als Konfidenzintervall erhalten wir

$$[\,3.1 - 2.50, 3.1 + 2.50\,] = [\,0.60, 5.60\,].$$

Das ist ziemlich groß, aber die Varianz ist ja auch erheblich.

C. *Mittelwertabstand zweier Normalverteilungen mit bekannter Streuung*

Problem: Gegeben seien zwei Größen, die $N(a_1, \sigma_1^2)$- bzw. $N(a_2, \sigma_2^2)$-verteilt sind; dabei sind a_1 und a_2 unbekannt, aber σ_1 und σ_2 sind bekannt. Die erste wird m-mal, die zweite n-mal abgefragt; die Ergebnisse bezeichnen wir mit $x_1, \ldots, x_m$ bzw. $y_1, \ldots, y_n$, die Stichprobenmittel mit $\overline{x}$ und $\overline{y}$.

Wir wollen mit diesen Informationen ein Konfidenzintervall für $a_1 - a_2$ zum Konfidenzniveau $1-\alpha$ finden. (Das tritt zum Beispiel dann auf, wenn man wissen möchte, um wieviel erfolgreicher die eine Düngemethode als die andere ist.)

Lösung: Bestimme – mit Hilfe der Tafel der Standard-Normalverteilung – ein Intervall $[\,-r, r\,]$, das unter $N(0,1)$ die Wahrscheinlichkeit $1-\alpha$ hat. Setze

$$d := \frac{\sqrt{mn}}{\sqrt{n\sigma_1^2 + m\sigma_2^2}}.$$

Das gesuchte Konfidenzintervall für $a_1 - a_2$ ist dann $[\,\overline{x} - \overline{y} - r/d, \overline{x} - \overline{y} + r/d\,]$.

Begründung: In der elementaren Stochastik hatten wir gezeigt: Sind X, Y unabängig und $N(b_1, \sigma_1^2)$- bzw. $N(b_2, \sigma_2^2)$-verteilt, so ist cX $N(cb_1, c^2\sigma_1^2)$-verteilt und $X + Y$ $N(b_1 + b_2, \sigma_1^2 + \sigma_2^2)$-verteilt. Damit folgt: Die Verteilung von

$$[\overline{x} - \overline{y} - (a_1 - a_2)]d$$

ist $N(0, 1)$. Der Rest ist dann klar.

Beispiel: $\alpha = 0.05$, $m = 100$, $n = 200$, $\sigma_1 = 1$ und $\sigma_2 = 2$. Angenommen, wir erhalten $\overline{x} - \overline{y} = 3$. Da in diesem Fall $d = \dfrac{\sqrt{20000}}{\sqrt{200 + 4 \cdot 100}} = 5.77$ ist und wir $r = 1.96$ wählen können, folgt: Ein Konfidenzintervall für $a_1 - a_2$ ist durch $[\, 3 - 1.96/5.77, 3 + 1.96/5.77\,] = [\, 2.66, 3.34\,]$ gegeben.

D. *Quotient der Varianzen zweier Normalverteilungen mit bekannten Mittelwerten*

Problem: Gegeben seien wieder zwei Größen, die $N(a_1, \sigma_1^2)$- bzw. $N(a_2, \sigma_2^2)$-verteilt sind; diesmal sind a_1 und a_2 bekannt, uns interessiert ein Konfidenzintervall der Form $[\, r_1, r_2\,]$ zur Irrtumswahrscheinlichkeit α von σ_2^2/σ_1^2. (Man denke an den Vergleich zweier Werkzeugmaschinen.) Die erste Größe wird m-mal, die zweite n-mal abgefragt; die Ergebnisse bezeichnen wir mit $x_1, \ldots, x_m$ bzw. $y_1, \ldots, y_n$.

Lösung: Berechne die Größe

$$q := \frac{\sum (x_i - a_1)^2/m}{\sum (y_i - a_2)^2/n}$$

und suche aus einer Tabelle ein Intervall $[\, s_1, s_2\,]$, so dass die Wahrscheinlichkeit für $[\, s_1, s_2\,]$ unter $F_{m,n}$ gleich $1 - \alpha$ ist. Setze dann $[\, r_1, r_2\,] := [\, s_1/q, s_2/q\,]$.

Begründung Die $w_i := (x_i - a_1)/\sigma_1$ und die $z_i := (y_i - a_2)/\sigma_2$ sind $N(0, 1)$-verteilt. Folglich ist

$$\frac{\sum w_i^2/m}{\sum z_i^2/n} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{\sum (x_i - a_1)^2/m}{\sum (y_i - a_2)^2/n} = \frac{\sigma_2^2}{\sigma_1^2} q$$

nach Satz 2.5.6 $F_{m,n}$-verteilt.

> Gesucht sind in der Regel die x und y mit $P(F_{m,n} \geq x) = \alpha/2$ und $P(F_{m,n} \leq y) = \alpha/2$. Manchmal sind nur die x-Werte in den Tabellen verzeichnet. Beachte dann, dass aufgrund der Definition $F_{m,n} \leq y$ genau dann gilt, wenn $F_{n,m} \geq 1/y$. Suche also x mit $P(F_{n,m} \geq x) = \alpha/2$ und setze dann $y := 1/x$.

Beispiel: Sei $n = 5, m = 7$, $\alpha = 0.1$. Angenommen, es ergibt sich $q = 10$. Aus der Tabelle entnehmen wir

$$[\, s_1, s_2\,] = [\, 0.205, 3.97\,],$$

und wir schließen: Unser Konfidenzintervall für σ_2^2/σ_1^2 ist $[\, 0.02, 0.39\,]$.

E. *Varianz einer Normalverteilung bei bekanntem Erwartungswert*

Problem: Eine Normalverteilung $N(a, \sigma^2)$ wird n-mal abgefragt (mit dem Ergebnis $x_1, \ldots, x_n$). a ist bekannt, ein Konfidenzintervall für σ^2 zum Niveau α ist gesucht.

Lösung: Bestimme ein Intervall $[\, a_1, a_2 \,]$, das unter χ^2_n die Wahrscheinlichkeit $1-\alpha$ hat und berechne $d := \sum (x_i - a)^2$. Dann liegt d/σ^2 mit Wahrscheinlichkeit $1-\alpha$ in $[\, a_1, a_2 \,]$, d.h. $[\, d/a_2, d/a_1 \,]$ ist ein Konfidenzintervall für σ^2.

Begründung Die $(x_i - a)/\sigma$ sind $N(0,1)$-verteilt, folglich ist $\sum (x_i - a)^2/\sigma^2$ nach Korollar 2.5.4 $\chi^2_n - verteilt$

Beispiel: Es gebe 10 Abfragen, und $\alpha = 0.05$ sei vorgegeben. Der Tabelle (10 Freiheitsgrade!) entnehmen wir, dass $[\, a_1, a_2 \,] = [\, 3.25, 20.48 \,]$ die geforderten Eigenschaften hat. Ist also etwa bei einer konkreten Messung das sich ergebende d gleich 123, so ergibt sich daraus das Konfidenzintervall

$$[\, 123/20.48, 123/3.25 \,] = [\, 6.00, 37.84 \,]$$

für σ^2.

F. *Varianz einer Normalverteilung bei unbekanntem Erwartungswert*

Problem: Wie eben, aber mit unbekanntem Erwartungswert.

Lösung: Bestimme ein Intervall $[\, a_1, a_2 \,]$, das unter χ^2_{n-1} die Wahrscheinlichkeit $1-\alpha$ hat und berechne $\tilde{d} := \sum (x_i - \overline{x})^2$. Dann liegt $\tilde{d}/\sigma^2$ mit Wahrscheinlichkeit $1-\alpha$ in $[\, a_1, a_2 \,]$, d.h. $\left[\, \tilde{d}/a_2, d/a_1 \,\right]$ ist ein Konfidenzintervall für σ^2.

Begründung Hier wird Satz 2.5.10(ii) angewendet.

Beispiel: Wir betrachten die Situation aus „E“, es sei aber a unbekannt. Diesmal müssen wir in der Tabelle bei 9 Freiheitsgraden nachschauen, es ergibt sich das Intervall

$$[\, a_1, a_2 \,] = [\, 2.70, 19.02 \,] .$$

Das impliziert – falls wir wieder $\tilde{d} = 123$ annehmen – das Konfidenzintervall

$$[\, 123/19.02, 123/2.70 \,] = [\, 6.47, 45.55 \,] .$$

Das ist erwartungsgemäß größer als das in „E“, die Verschlechterung ist auf die fehlende Information über a zurückzuführen.

Kapitel 3

Testtheorie

Bisher ging es darum, gewisse aus dem Wahrscheinlichkeitsmaß $\mathbb{P}_\theta$ abgeleitete Parameter $\gamma(\theta)$ möglichst gut zu schätzen. In diesem Kapitel wollen wir *Entscheidungen aufgrund vorliegender statistischer Erhebungen* treffen. Dazu wird zunächst in *Abschnitt 3.1* der allgemeine Rahmen präzisiert: Was ist eine Hypothese, was ist die Aufgabe des Statistikers, welche Fehler kann man machen? Anschließend wollen wir in *Abschnitt 3.2* detailliert untersuchen, wie man im Fall von zwei Wahrscheinlichkeitsräumen mit Hilfe des Konzepts der zufälligen Testfunktion eine optimale Lösung des Testproblems finden kann. In *Abschnitt 3.3* sollen diese Resultate dann auf allgemeinere Fragestellungen angewendet werden, auch sollen die im zweiten Kapitel entwickelten Methoden für die Behandlung des Testproblems eingesetzt werden.

3.1 Hypothesen

Im wirklichen Leben sind permanent irgendwelche Entscheidungen zu treffen, die meisten treffen wir sicher unbewusst. Nehmen wir zum Beispiel die Frage, ob Sie heute einen *Schirm mitnehmen* sollten, wenn Sie das Haus verlassen. Grundlage Ihrer Entscheidung wird doch eine Einschätzung sein: Wird es heute regnen oder nicht?

Mal angenommen, Sie tippen auf Regen und nehmen den Schirm mit. Wenn es dann wirklich regnet, dann haben Sie richtig vorgesorgt. Andernfalls tragen Sie Ihren Schirm den ganzen Tag lang völlig nutzlos spazieren.

Sind Sie dagegen optimistisch und rechnen mit gutem Wetter, so werden Sie bei dem ersten Regenschauer ein Problem haben.

Das Fazit: Wenn man die Wahl zwischen zwei Alternativen hat, so kann sich diese Wahl nachträglich als günstig oder ungünstig erweisen, je nachdem, was denn nun wirklich passiert.

In der Statistik hat sich die Terminologie eingebürgert, eine Annahme über den wirklichen Zustand der Welt eine *Hypothese* zu nennen, genauer spricht

man von einer *Nullhypothese*, das Gegenteil heißt die *Alternativhypothese*. (In unserem Beispiel könnte man „Es bleibt trocken“ oder aber auch „Es wird regnen“ als Nullhypothese deklarieren. Die Alternativhypothese ist dann die jeweils andere Aussage.)

Hier einige typische Beispiele, die in der Statistik eine Rolle spielen:

1. Die Nullhypothese könnte lauten: Dies ist ein fairer Würfel. Die Alternativhypothese wäre dann: Der Würfel ist unfair.

2. Gegeben sei das aus allen Normalverteilungen bestehende statistische Modell. Eine mögliche Nullhypothese wäre: Der Mittelwert μ ist gleich 12.

3. Wie vorstehend, aber mit der Nullhypothese $\mu \leq 12$.

Es ist nicht schwer, sich weitere Beispiele mit statistischen Modellen und aus dem wirklichen Leben auszudenken. Es gibt aber nun ein Problem mit der Einschätzung des Wahrheitsgehalts von Hypothesen:

Man kann Fehler machen!

Mal angenommen, wir nennen die Nullhypothese H_0. Wenn wir glauben, dass sie falsch ist und uns danach richten, können zwei Dinge passieren. Sie kann wirklich falsch sein, dann ist alles in Ordnung. Sie kann aber auch wahr sein, dann haben wir einen Fehler gemacht.

Wir könnten aber auch glauben, dass H_0 wahr ist. Falls ja, ist wieder alles bestens, andernfalls haben wir falsch getippt und müssen eventuell mit Konsequenzen rechnen.

Definition 3.1.1. *Es seien Nullhypothese H_0 und Alternativhypothese H_1 gegeben.*

(i) Nimmt man an, dass H_0 nicht *gilt, obwohl H_0 in Wirklichkeit gilt, so nennt man das einen* Fehler erster Art.

(ii) Nimmt man an, dass H_1 nicht *gilt, obwohl H_1 in Wirklichkeit gilt, so nennt man das einen* Fehler zweiter Art.

Bemerkungen und Beispiele:

1. Das ist wirklich etwas verwirrend. Daher soll die Definition noch einmal in einer kleinen Tabelle wiederholt werden, die man immer vor Augen haben sollte:

	H_0 angenommen	H_1 angenommen
H_0 wirklich	o.k.!	Fehler 1. Art
H_1 wirklich	Fehler 2. Art	o.k.!

2. Es ist im Grunde völlig beliebig, welche der Hypothesen H_0 und welche H_1 ist. Es hat sich aber die folgende *Konvention* eingebürgert:

Die Hypothesen sind so zu bezeichnen, dass der Fehler erster Art schwerwiegendere Konsequenzen als der Fehler zweiter Art hat.

Typischerweise wird das folgende Beispiel in der Literatur angeboten. Man stellt sich eine Feuerwehrzentrale vor, soeben ist ein Anruf eingegangen: „Das mathematische Institut brennt!“

- H_0: Der Anrufer hat Recht, es brennt wirklich.

- H_1: Es war ein Witzbold, der Anruf war eine Folge seiner Prüfungsangst.

Der Fehler erster Art bestünde in diesem Beispiel darin anzunehmen, dass es nicht brennt und genervt aufzulegen, obwohl alles bereits in Flammen steht. Und einen Fehler zweiter Art würde die Feuerwehr begehen, wenn sie ausrückt und am Ziel nur überraschte Mathematiker findet.

Es ist klar, dass der Fehler erster Art in diesem Beispiel als der gravierendere einzuschätzen wäre, deswegen wäre die Bezeichnungsweise im Einklang mit der Konvention.

Es sollte allerdings betont werden, dass das in vielen Fällen durchaus nicht so klar ist, dass also unterschiedliche Betrachter durchaus verschiedener Meinung sein können, welcher der Fehler gravierender ist.

> *Ein Beispiel dazu:* Die Klausuraufgaben für die bundesweite Mathematik-Vordiplomprüfung werden mit der Post verschickt.
>
> Hypothese: Der Postweg ist sicher und zuverlässig.
>
> Ein möglicher Fehler wäre: Die Post ist wirklich zuverlässig, man beauftragt aber trotzdem einen Kurierdienst. Das würde der Finanzsenator als schwerwiegend ansehen. Variante: Man vertraut der Post, die Klausur geht aber verloren. Dann hätten die Prüfungsbüros ein Problem.

Es hilft aber nichts, man muss sich entscheiden, denn irgendwie müssen die möglichen Hypothesen ja bezeichnet werden.

3. Hier einige Beispiele für Nullhypothesen aus dem täglichen Leben. Versuchen Sie erstens, in den einzelnen Fällen die Fehler erster Art und zweiter Art verbal zu beschreiben und beobachten Sie sich zweitens im Alltag, wann Sie – mehr oder weniger unbewusst – mit dem Abwägen dieser beiden Fehlerarten beschäftigt sind.

- Die Kandidatin für die Vordiplomsprüfung stellt die Hypothese auf, dass die Fragen bei dem Prüfer B. wohl ganz harmlos sein werden.

- Herr R. kommt in die Disco und sieht eine schöne Frau, die er gerne ansprechen würde. Neben ihr steht allerdings ein ziemlich brutal aussehender männlicher Mensch. Seine Hypothese: Es ist ihr Bruder.

- Herr B. möchte einen Badezimmerschrank anbringen und muss ein Loch bohren. Seine Hypothese: Das Stromkabel läuft ganz woanders.

Es ist nun *Zeit, etwas formaler zu werden*. Der passende Rahmen ist das Zugrundelegen eines statistischen Modells, d.h. einer Familie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ von Wahrscheinlichkeitsmaßen, die alle auf dem gleichen Messraum $(\Omega, \mathcal{E})$ definiert sind. Wir nehmen an, dass Θ disjunkt in zwei nicht leere Teilmengen Θ_0 und Θ_1 zerlegt ist, dann entspricht die Nullhypothese der Aussage $\theta \in \Theta_0$ und die Alternativhypothese der Aussage $\theta \in \Theta_1$. Auf diese Weise können nicht alle der vorstehenden Beispiele formalisiert werden, wir werden uns auf Hypothesen des Typs „Die Erfolgswahrscheinlichkeit ist ≤ 0.2“ oder „Der Mittelwert der normalverteilten Zufallsvariable ist 66“ beschränken.

Wie in der Schätztheorie wird ein θ ausgewählt, und aus $(\Omega, \mathcal{E}, \mathbb{P}_\theta)$ werden n unabhängige Abfragen gezogen. Aufgrund des Ergebnisses $(x_1, \ldots, x_n) \in \Omega^n$ sollen wir uns für H_0 oder H_1 entscheiden, d.h. es ist eine Abbildung

$$d : \Omega^n \to \{0, 1\}$$

anzugeben. (Interpretation: $d(x) = i$ bedeutet, dass wir uns für H_i entscheiden[1].) So ein d heißt dann eine *Testfunktion* (auch einfach „Test“ oder „Entscheidungsfunktion“).

Im Zusammenhang mit Tests spielen *zwei Begriffe* eine wichtige Rolle

- Die durch $G_d : \theta \mapsto E_\theta d = \mathbb{P}^n_\theta(\{d = 1\})$ von Θ nach $\mathbb{R}$ hat dann die folgende Interpretation:
  - Für $\theta \in \Theta_0$ ist $G_d(\theta)$ die Wahrscheinlichkeit für einen Fehler erster Art. (Beachte nur: $G_d(\theta)$ ist das Maß der Menge $\{d = 1\}$ unter $\mathbb{P}^n_\theta$, also die Wahrscheinlichkeit, fälschlich für H_1 zu votieren.)
  - Für $\theta \in \Theta_1$ ist $1 - G_d(\theta)$ die Wahrscheinlichkeit für einen Fehler zweiter Art. (Denn diese Zahl ist die Wahrscheinlichkeit von $\{d = 0\}$ unter $\mathbb{P}^n_\theta$.)

  G_d heißt *die Gütefunktion* des Tests d.
- Die Menge $d^{-1}(\{1\})$ heißt der *kritische Bereich* (auch: *Verwerfungsbereich* oder *Ablehnungsbereich*) des Tests.

Als Fazit erhalten wir:

> Im Idealfall findet man ein d, so dass die Gütefunktion auf Θ_0 nahe bei Null und auf Θ_1 nahe bei Eins ist.

So etwas gibt es im Allgemeinen nicht, und deswegen muss man nach sinnvollen Kompromissen suchen. Je nach Zielvorgabe kann das Ergebnis ganz unterschiedlich ausfallen, mögliche Strategien sind:

[1] Man ist in der Statistik mit den Formulierungen etwas vorsichtiger. Statt „H_0 wird angenommen“ sagt man „H_0 wird nicht abgelehnt“, und statt „Entscheidung für H_1“ sagt man „H_0 wird abgelehnt“. Das trifft das, was wirklich ausgesagt werden kann, auch besser: Wenn eine Münze auf Fairness getestet wird, so ist eine Aussage „Die Hypothese, dass die Münze fair ist, kann nicht abgelehnt werden“ angemessener als eine der Form „Diese Münze ist aufgrund des Tests als fair anzusehen“.

1. Suche ein d so, dass die Summe aus den maximalen Fehlern erster und zweiter Art möglichst klein ist. (So verfährt man wohl manchmal bei Alltagsentscheidungen.)

2. Fixiere ein $\alpha > 0$ und beschränke dich dann auf die d, für die alle Fehler erster Art durch α beschränkt sind. Suche unter diesen d eines, für das der maximale Fehler zweiter Art so klein wie möglich ist. Dieser Ansatz ist sicher dann gerechtfertigt, wenn man – wie im Feuerwehrbeispiel – Fehler erster Art unbedingt vermeiden möchte.

Definition 3.1.2. *Sei $\alpha > 0$. EinTest d heißt* Test zum Irrtumsniveau α *(oder zum Konfidenzniveau $1 - \alpha$), wenn $G_d(\theta) \leq \alpha$ für alle $\theta \in \Theta_0$ gilt.*

Zur Illustration der neu eingeführten Begriffe wird jetzt *ein einfaches Beispiel* ausführlich diskutiert. Hier die Ausgangssituation:

> Im Baumarkt gibt es schon im Januar Tulpenzwiebeln, zwei Mischungen sind im Angebot. Bei Variante 1 kauft man 10 rote Tulpen, 20 weiße Tulpen und 70 gelbe Tulpen, bei Variante 2 sind es dagegen 55 rote, 35 weiße und 10 gelbe. Leider sieht man den Zwiebeln die Farbe der zukünftigen Blüte nicht an.
>
> Sie haben eine Packung gekauft, sich aber nicht gemerkt, aus welchem Korb Sie die genommen haben. Und der Zettel mit der Beschreibung ist unterwegs verloren gegangen.
>
> Ihre Nullhypothese lautet: „Es handelt sich um Variante 1“. Zum Test bringen Sie eine Zwiebel im Schnelldurchgang im Gewächshaus zur Blüte. Welche Testfunktion sollte man sinnvollerweise auswählen?

Qualitativ ist zum Beispiel offensichtlich, dass das Ergebnis „gelb“ eher für die Nullhypothese spricht als das Ergebnis „rot“. Um alles quantitativ behandeln zu können, stellen wir alle Testfunktionen systematisch zusammen. Dazu reicht es, den kritischen Bereich (auf dem $d = 1$ ist, in dem die Nullhypothese also abgelehnt wird) anzugeben. Nachstehend sind alle Möglichkeiten für kritische Bereiche zusammen mit den Wahrscheinlichkeiten für Fehler erster und zweiter Art[2)] zusammengestellt. Bei uns ist $\Omega = \{\text{rot}, \text{weiß}, \text{gelb}\}$, und n ist gleich 1, da wir nur eine Blütenfarbe testen. Folglich gibt es $2^3 = 8$ Tests:

Nummer	kritischer Bereich	W. für Fehler 1. Art	W. für Fehler 2. Art
1	$\emptyset$	0	1
2	{rot}	0.1	0.45
3	{weiß}	0.2	0.65
4	{gelb}	0.7	0.9
5	{rot,weiß}	0.3	0.1
6	{rot,gelb}	0.8	0.35
7	{weiß, gelb}	0.9	0.55
8	{rot, weiß, gelb}	1	0.

[2)] Also den Zahlen $G_d(\theta)$ für die Fälle θ=„Variante 1“ bzw. θ=„Variante 2“.

Hier wird noch einmal klar, dass die Summe der Wahrscheinlichkeiten für die Fehler erster und zweiter im Allgemeinen nicht 1 ist: Es werden zwar Wahrscheinlichkeiten für disjunkte Ereignisse addiert (kritischer Bereich und Komplement), die Wahrscheinlichkeitsmaße sind aber verschieden.

Nun können Tests mit verschiedenen Eigenschaften leicht bestimmt werden. Zum Beispiel:

- Bei Test 1 macht man garantiert keinen Fehler erster Art, dafür handelt man sich einen maximalen Fehler zweiter Art ein.
- Unter allen Tests vom Irrtumsniveau ≤ 0.2 ist Test 2 der beste, denn bei ihm ist die Wahrscheinlichkeit für einen Fehler zweiter Art minimal. (Das ist auch plausibel: Bei einer roten Blüte ist kaum zu erwarten, dass Variante 2 vorlag.)

Wir werden uns hauptsächlich um Tests zu einem vorgegebenen Konfidenzniveau kümmern und versuchen, unter diesen Kandidaten einen bestmöglichen zu finden (also die Wahrscheinlichkeit für den Fehler zweiter Art zu minimieren). Bevor wir damit im nächsten Abschnitt beginnen, sollen noch kurz einige andere Ansätze besprochen werden.

Verlustmatrix

Je nachdem, wie man sich entscheidet, wird man mit einer von vier Möglichkeiten rechnen müssen: Man entscheidet sich für H_0 oder H_1, und tatsächlich kann H_0 oder H_1 gelten. Mit jeder dieser vier Möglichkeiten ist eine gewisse Wichtung verbunden, man versucht, diese durch eine Zahl zu beschreiben. Das führt zur Definition der *Verlustmatrix:*

> Bezeichne mit L_{ij} die Bewertung des Verlustes, wenn in Wirklichkeit H_i gilt und man sich für H_j entscheidet $(i, j = 0, 1)$. Die 2×2-Matrix (L_{ij}) heißt dann die *Verlustmatrix*.
>
> Die L_{ij} dürfen *auch negativ* sein, das kommt zum Beispiel dann vor, wenn eine richtige Entscheidung einen Gewinn bringt.

Abstrakt sieht es also so aus (L_{10} etwa bezeichnet den Verlust bei einem Fehler zweiter Art):

	H_0 angenommen	H_1 angenommen
H_0 wirklich	L_{00}	L_{01}
H_1 wirklich	L_{10}	L_{11}

Je nach Situation kann die Verlustmatrix sehr unterschiedlich aussehen, im Feuerwehrbeispiel etwa könnte man vielleicht mit den folgenden Zahlen (in Euro) arbeiten:

	H_0 angenommen	H_1 angenommen
H_0 wirklich	5000	10 000 000
H_1 wirklich	1000	0

Hat man sich auf eine Verlustmatrix geeinigt, kann man die Güte eines Tests auf sehr unterschiedliche Weise bewerten, es folgen die wichtigsten Beispiele.

Risikofunktion

Wir wollen für den Rest des Abschnitts annehmen, dass Θ_0 und Θ_1 einelementig sind: $\Theta_0 = \{\mathbb{P}_0\}$, $\Theta_1 = \{\mathbb{P}_1\}$.

Es sei $d : \Omega \to \{0, 1\}$ ein Test. Bezeichne für $i, j = 0, 1$ mit α_{ij} die Wahrscheinlichkeit, dass – wenn wirklich H_i zutrifft – der Test das Ergebnis j liefert. (Zum Beispiel ist α_{01} die Wahrscheinlichkeit für einen Fehler 1. Art.) Unter dem zu d gehörigen *Risiko* versteht man dann die Zahlen

$$R_d(i) := \sum_{j=0,1} L_{ij}\alpha_{ij};\ i = 0, 1;$$

damit ist $R_d(i)$ so etwas wie der Erwartungswert des Verlusts, wenn H_i zutrifft.

Es ist klar, dass man ein d sucht, für das das Risiko „möglichst klein“ ist. Doch was soll das bedeuten? Durchläuft d alle möglichen Testfunktionen – für ein k-elementiges Ω sind das 2^k Funktionen – so geben die Tupel $(R_d(0), R_d(1))$ zu einer „Punktwolke“ W im $\mathbb{R}^2$ Anlass. Und es ist durchaus nicht klar, welcher Punkt von einem besonders „günstigen“ d stammt. Wir diskutieren zwei mögliche Lösungsansätze.

Bayes-Ansatz

Beim Bayes-Ansatz nimmt man eine a-priori-Verteilung für das Auftreten von H_0 und H_1 an. Gegeben sind also zwei nichtnegative Zahlen α_0 und α_1 mit $\alpha_0 + \alpha_1 = 1$, wobei α_i die geschätzte Wahrscheinlichkeit für das Eintreten von H_i ist. Geht man von diesen Wahrscheinlichkeiten aus, so ist

$$V_d := \alpha_0 R_d(0) + \alpha_1 R_d(1)$$

der Erwartungswert des Verlusts, wenn man mit dem Test d arbeitet.

Um zu entscheiden, welches d optimal ist, muss nur an die elementare geometrische Tatsache erinnert werden, dass die Menge aller (x, y), für die $\alpha_0 x + \alpha_1 y = c$ gilt, eine Gerade ist und dass diese Geraden für verschiedenes c parallel sind. Um eine optimale Lösung zu finden, muss man also das c in dieser Geradenschar so klein wie möglich wählen, dass es gerade noch die Punktwolke W trifft. Die d, für die $(R_d(0), R_d(1))$ zu den Schnittpunkten gehört, sind dann bestmöglich.

Minimax-Lösungen

Traut man den a-priori-Wahrscheinlichkeiten nicht, kann man auf eine sehr konservative Strategie ausweichen: Man versucht, das d so zu wählen, dass das

Maximum der Verluste unter Kontrolle bleibt. Genauer: Man betrachtet die Zahlen

$$\max\{R_d(0), R_d(1)\}$$

und bestimmt d so, dass diese Zahl dafür minimal ist. (Man spricht von einer *Minimax-Lösung*.)

Auch das kann man sich veranschaulichen: Man fixiert zunächst ein „sehr kleines“ r und betrachtet alle (x, y) mit $\max\{x, y\} = r$; diese Menge soll Δ_r heißen.

Lässt man nun r wachsen, so wird Δ_r irgendwann einmal erstmalig die Punktwolke W berühren. Alle d, die zu getroffenen $(R_d(0), R_d(1))$ gehören, sind Minimax-Lösungen.

Zum Schluss dieses Abschnitts soll noch kurz *eine weitere Vokabel* erläutert werden: Was ist der *p-Wert eines Tests*? Angenommen, wir testen H_0 gegen H_1, zur Verfügung stehen verlässliche Tests d_α zum Niveau α für alle α. Wir fixieren nun α_0 und nehmen eine Stichprobe $x_1, \ldots, x_n$. Es sei so, dass unser Test d_{α_0} aufgrund dieser Stichprobe sagt: H_0 wird nicht abgelehnt.

Nun kann es sein, dass auch für gewisse $\alpha > \alpha_0$ die Hypothese bei diesem konkreten Ergebnis nicht abgelehnt worden wäre. Das ist dann eine Verstärkung der Empfehlung, nicht abzulehnen, denn der Ablehnungsbereich ist nun größer geworden.

> Ein Beispiel: Die Hypothese lautet $p \leq 0.5$, sie soll zum Niveau α getestet werden, und es gebe bei n Versuchen k Erfolge. Wenn dann k „nicht zu groß“ ist, wird sie nicht abgelehnt. Falls nun sogar k „sehr klein“ ist, sind wir uns ganz sicher, dass H_0 gilt.

Quantifiziert wird das durch die folgende Definition: Der *p-Wert* eines Tests[3)] ist das Supremum der Zahlen α, für die H_0 nicht abgelehnt worden wäre. Ein großer p-Wert kann also so interpretiert werden, dass die Empfehlung „H_0 nicht ablehnen!“ besonders vertrauenserweckend ist.

> Achtung: Vorher wurden „Test“ und „Testfunktion“ synonym gebraucht, hier beim p-Wert geht es wirklich um konkretes Erzeugen von Stichproben. Anders ausgedrückt: Wenn eine Testfunktion d gewählt ist, wird es bei verschiedenen Abfragen in der Regel auch verschiedene p-Werte geben.

Statistik-Lehrbücher weisen in diesem Zusammenhang gern darauf hin, dass ein großer p-Wert nicht dazu verführen darf, sich in dem Glauben zu wiegen, dass man einen Test zu einem sehr hohen Konfidenzniveau durchgeführt hat: Das α ist unbedingt vorher festzusetzen, nach dem Test darf es nicht mehr verändert werden. Hiermit wird diese Warnung weitergegeben.

[3)] Genauer: Einer Familie von Tests.

3.2 Tests mit vorgegebenem Konfidenzniveau: Alternativtests

In diesem Abschnitt soll gezeigt werden, wie man einen besten stochastischen Test im Fall einer Alternativentscheidung finden kann. Wir gehen von einer Situation aus, in der das statistische Modell aus nur zwei Wahrscheinlichkeitsmaßen $\mathbb{P}_0$ und $\mathbb{P}_1$ besteht. Es wird $\mathbb{P}$ aus $\Theta := \{\mathbb{P}_0, \mathbb{P}_1\}$ gewählt, und wir sollen aufgrund einer Stichprobe zum Niveau α entscheiden, ob die Nullhypothese $\mathbb{P} = \mathbb{P}_0$ abgelehnt werden sollte. Ohne Einschränkung bestehe die Stichprobe nur aus einem Element, das lässt sich durch Übergang zum n-fachen Produkt stets erreichen. Wir behandeln zunächst den diskreten Fall und diskutieren dann die Modifikationen, die im Fall von Maßen mit Dichten notwendig sind.

Der Fall diskreter Räume

α sei vorgelegt. Mal angenommen, wir wollen einen Test zur Irrtumswahrscheinlichkeit α konstruieren. Dann müssen wir doch soviele x zu einer Menge K[4] zusammenfassen (um dann den Test d als charakteristische Funktion von K zu definieren), dass $\mathbb{P}_0(K)$ möglichst genau gleich α ist: *So* wird garantiert, dass der Fehler erster Art α nicht übersteigt. Möchte man den Fehler zweiter Art gleichzeitig möglichst klein halten, sollte $\mathbb{P}_1(K)$ „möglichst groß“ sein. Kurz: Es ist naheliegend, K aus solchen x zusammenzusetzen, für die $\mathbb{P}_0$ klein und $\mathbb{P}_1$ groß ist.

Es kann dadurch ein kleines Problem geben, dass man bei diesem „Einsammeln“ nicht genau den Wert α trifft. Das wird auf elegante Weise dadurch gelöst, dass man auch *stochastische Entscheidungen* zulässt. Ab hier betrachten wir also statt $d : \Omega^n \to \{0, 1\}$ Funktionen $d : \Omega^n \to [\,0, 1\,]$. Das wird so interpretiert: Ist $d(x) = p$, entscheide dich mit Wahrscheinlichkeit p für H_1. Es ist klar, dass die bisherigen Tests als Spezialfälle enthalten sind und dass der Erwartungswert von d wieder die Wahrscheinlichkeit für einen Fehler erster Art ist.
Ein derartiges d heißt ein *stochastischer Test*.

Es ist Zeit für eine formalere

Definition 3.2.1. *Mit den vorstehenden Bezeichnungen definieren wir*

$$R(x) := \frac{\mathbb{P}_1(\{x\})}{\mathbb{P}_0(\{x\})}$$

für $x \in \Omega$[5]

Ein stochastischer Test d heißt ein Neyman-Pearson-Test *zum Niveau α, falls gilt:*

(i) Der Erwartungswert von d unter $\mathbb{P}_0$ ist α.

[4] K heißt auch der *kritische Bereich*.

[5] Wir dürfen voraussetzen, dass es keine Quotienten der Form $0/0$ gibt; ansonsten gelten die Rechenregeln in der Kompaktifizierung von $\mathbb{R}$.

(ii) Es gibt eine Zahl c, so dass $d(x) = 1$ (bzw. gleich 0) ist für $R(x) > c$ (bzw. $< c$). Auf der Menge $\{R = c\}$ darf d beliebige Werte haben.

Satz 3.2.2. *Es gibt einen Neyman-Pearson-Test zum Niveau α, und jeder derartige Test ist bestmöglich im folgenden Sinn: Andere stochastische Tests zum Niveau α führen zu einem nicht kleineren Fehler zweiter Art. Außerdem gilt: Jeder beste Test ist ein Neyman-Pearson-Test.*

Beweis: Wir stellen uns die Elemente von $\Omega = \{x_1, \ldots\}$ so sortiert vor, dass R monoton fällt. Wir wählen dann ein c, so dass $\mathbb{P}_0(\{x \mid R(x) > c\}) < \alpha$ und $\mathbb{P}_0(\{x \mid R(x) \geq c\}) \geq \alpha$; es ist klar, dass so ein c existiert.

Definiere $\Delta := \{R = c\}$. d ist durch die Werte 1 (auf $\{R > c\}$), 0 (auf $\{R < c\}$) und γ (auf Δ) definiert. Dabei ist

$$\gamma := \frac{\alpha - \mathbb{P}_0(\{R > c\})}{\mathbb{P}_0(\Delta)}.$$

Es ist dann klar, dass d ein Neyman-Pearson-Test zum Niveau α ist.

Sei nun ψ ein beliebiger Test zum Niveau α und d ein entsprechender Neyman-Pearson-Test. Wir wollen zeigen, dass d nicht schlechter als ψ ist.

> Nebenbei sei auf die *geometrische Deutung* dieses Sachverhalts hingewiesen:
>
> Gegeben seien zwei Konvexkombinationen $\lambda_1, \ldots, \lambda_n$ und $\mu_1, \ldots, \mu_n$. Versuche, Zahlen $a_1, \ldots, a_n \in [\,0, 1\,]$ so zu finden, dass $\sum a_i \lambda_i = \alpha$ ist und $\sum a_i \mu_i$ so groß wie möglich ist. Es ist zu zeigen, dass die optimale a_i-Wahl darin besteht, sie auf den i mit großem μ_i/λ_i gleich Eins zu setzen, auf den i mit kleinem μ_i/λ_i gleich Null und zwischendurch evtl. einmal in $[\,0, 1\,]$, um $\sum a_i \lambda_i = \alpha$ zu erreichen.

Setze $a_i := \psi(x_i)$. Mal angenommen, es wäre $a_1 < 1$. Suche ein j mit $a_j > 0$ und betrachte für „kleines“ $\varepsilon > 0$ einen Test ψ': Der ist auf allen $i \neq 1, j$ wie ψ definiert, bei 1 hat er den Wert $a_1 + \varepsilon$ und bei j den Wert $a_j - \varepsilon(\mathbb{P}_0(x_1)/\mathbb{P}_0(x_j))$. Auch ψ' ist ein Test zum Niveau α, aber der Fehler zweiter Art ist eher kleiner geworden.

Kurz: Man kann von ψ zu einem nicht schlechteren Test übergehen, für den $a_1 = 1$ ist. Genau so zeigt man, dass man durch Verbesserung $a_2 = 1$, $a_3 = 1$ usw. erreichen kann, so lange, bis das $\mathbb{P}_0$-Maß von $\{x_1, \ldots, x_i\}$ den Wert α nicht übersteigt.

Auf diese Weise folgt, dass man einen beliebigen Test zum Niveau α durch Übergang zu einem Neymann-Pearson-Test verbessern kann, und damit ist der Satz vollständig bewiesen □

Ein Beispiel:

1. p_0 und p_1 seien zwei Zahlen in $]\,0, 1\,[$, es gelte $p_0 < p_1$; man denke an eine Münze, die aus einer von zwei Produktionslinien für gefälschte Münzen stammt,

die Wahrscheinlichkeit für „Kopf“ kann p_0 oder p_1 sein. Die Münze wird nun n-mal geworfen, aus der Anzahl k der Erfolge soll man sich bei vorgegebenem Konfidenzniveau $1-\alpha$ für $H_0 : p = p_0$ oder $H_1 : p = p_1$ entscheiden.

Es geht also um zwei Binomialverteilungen auf $\{0, \ldots, n\}$, die erste hat ihren „Buckel“ links von dem der zweiten. Praktischerweise sind die Quotienten schon sortiert, wir behaupten nämlich:

$$\text{Die Abbildung } k \mapsto R(k) := \frac{b(k, n; p_1)}{b(k, n; p_0)} \text{ ist monoton steigend.}$$

(Zum Beweis betrachte man den Ausdruck $R(k+1)/R(k)$. Er ist – wie man leicht durch Umstellen zeigt – genau dann > 1, wenn $p_0 < p_1$ gilt.) Folglich müssen wir ein k_0 suchen, so dass

- $\sum_{k=k_0}^{n} b(k, n; p_0) \leq \alpha$.
- k_0 ist kleinstmöglich mit dieser Eigenschaft.

Es ist dann $\eta := \sum_{k=k_0}^{n} b(k, n; p_0) \leq \alpha$. Definiere d wie folgt:

- Für $k < k_0 - 1$ ist $d(k) := 0$.
- Für $k \geq k_0$ ist $d(k) := 1$.
- $d(k_0 - 1)$ hat den Wert $(\alpha - \eta)/b(k_0 - 1, n; p_0)$.

Dann ist d ein Neyman-Pearson-Test zum Niveau α: Der Erwartungswert unter $\mathbb{P}_0$ von d ist wirklich

$$\begin{aligned} \sum_k d(k) b(k, n; p_0) &= d(k_0 - 1) b(k_0 - 1, n, p_0) + \sum_{k \geq k_0} d(k) b(k, n; p_0) \\ &= (\alpha - \eta) + \eta \\ &= \alpha. \end{aligned}$$

Aus dem Satz folgt, dass dieser Test den kleinsten Fehler zweiter Art realisiert.

Man beachte also: Werden k Erfolge erzielt und ist k „klein“ bzw. „groß“ , so entscheidet man sich für H_0 (bzw. H_1). Es gibt aber einen Grenzfall, nämlich $k = k_0 - 1$, da muss dann eine Zufallsentscheidung herangezogen werden.

> Hier ist ein konkretes Beispiel mit $p_0 = 0.5$ und einem beliebigen $p_1 > 0.5$.
>
> Zum Niveau $\alpha = 0.125$ erhält man dann bei 4 Versuchen den folgenden Neyman-Pearson-Test d (mit k bezeichnen wir die Erfolgsanzahl):
>
> - Für $k \leq 2$ ist $d(k) = 0$: Man bleibt dann bei H_0.
> - $d(3) = 0.25$: Es wird noch einmal ein Zufallsexperiment gestartet, das mit Wahrscheinlichkeit 0.25 eine 1 liefert. Nur dann sagen wir H_1, sonst H_0.

- $d(4) = 1$: Bei 4 Erfolgen wird H_0 abgelehnt.

Man rechnet leicht nach, dass d wirklich exakt das Niveau α hat.

Maße mit Dichten

Wir betrachten nun den Fall, dass Ω ein Intervall in $\mathbb{R}$ ist und die Maße durch zwei Dichten f_0 und f_1 gegeben sind. Diesmal ist die Funktion

$$x \mapsto R(x) := \frac{f_1(x)}{f_0(x)}$$

zu betrachten. Die Definition eines Neyman-Pearson-Tests ist analog zum diskreten Fall: d heißt *Neyman-Pearson-Test* zum Niveau α, falls d unter $\mathbb{P}_0$ den Erwartungswert α hat und es ein c so gibt, dass $d(x) = 1$ (bzw. gleich 0) ist für $R(x) > c$ (bzw. $< c$); in der Regel wird $\{R = c\}$ eine Nullmenge sein, so dass es sich um einen deterministischen Test handelt.

Die Aussage des vorigen Satzes 3.2.2 gilt dann analog: Beste Tests sind Neyman-Pearson-Tests und umgekehrt. (Der Beweis im kontinuierlichen Fall kann ähnlich wie im diskreten Fall geführt werden; vgl. das Buch von Georgii, Seite 251.)

Ein Beispiel:

Gegeben seien zwei Normalverteilungen $N(a_0, 1)$ und $N(a_1, 1)$ mit $a_0 < a_1$. Der Quotient der Dichtefunktionen, also

$$x \mapsto R(x) := \frac{\exp(-(x-a_1)^2/2)}{\exp(-(x-a_0)^2/2)}$$

stimmt bis auf einen positiven Faktor mit der monoton steigenden Funktion $x \mapsto \exp((a_1 - a_0)x)$ überein. Folglich gibt es ein $b \in \mathbb{R}$, so dass der Neyman-Pearson-Test d die charakteristische Funktion von $[\, b, \infty\, [$ ist.

Potenzreihenfamilien

In Definition 2.3.7 hatten wir Potenzreihenfamilien eingeführt: Das sind statistische Modelle, bei denen das Maß $\mathbb{P}_\theta$ für ein $\{x\}$ durch

$$c_\theta \theta^{t(x)} h(x)$$

gegeben ist, wobei t eine geeignete Funktion ist. (Im kontinuierlichen Fall müssen die Dichtefunktionen so aussehen.) Zu Stichproben mit n Ergebnissen gehören dann die Wahrscheinlichkeiten

$$(c_\theta)^n \theta^{t(x_1)+t(x_2)+\cdots+t(x_n)} h(x_1) \cdots h(x_n).$$

Es seien nun zwei Elemente θ_0 und θ_1 aus so einer Potenzreihenfamilie vorgelegt, es sei etwa $\theta_0 < \theta_1$. In diesem Fall ist die Funktion R durch

$$R : x \mapsto \frac{c_{\theta_1}^n}{c_{\theta_0}^n} \Big(\frac{\theta_1}{\theta_0}\Big)^{T(x)}$$

gegeben, wobei wieder $T(x) = T(x_1, \ldots, x_n) := t(x_1) + \cdots + t(x_n)$. R ist damit eine in $T(x)$ steigende Funktion, und es folgt, dass Neyman-Pearson-Tests die folgende Form haben müssen:

- $d(x) = 1$, falls $T(x) > c$.
- $d(x) = 0$, falls $T(x) < c$.
- Auf $\{T = c\}$ ist der Wert von d dadurch festgelegt, dass die Wahrscheinlichkeit für den Fehler erster Art gleich α ist.

Dieses Prinzip lag den beiden vorstehend beschriebenen Beispielen zugrunde.

3.3 Ein- und zweiseitige Tests, Tests im Zusammenhang mit der Normalverteilung

Alternativtests beschreiben nur eine recht spezielle Situation. Realistischer sind doch Probleme, bei denen die Hypothesen eine ganze Klasse von Elementen enthalten:

- H_0 : Die Erfolgswahrscheinlichkeit ist ≤ 0.4.
- H_0 : Der Erwartungswert dieser Normalverteilung ist ≥ 2121.
- H_0 : Der Erwartungswert dieser Normalverteilung ist ≤ 0.
- H_0 : Der Erwartungswert dieser Normalverteilung ist gleich 1.

In diesem Abschnitt wollen wir die Ergebnisse vom Neyman-Pearson-Typ auf solche Fragen anwenden.

Einseitige Tests

Hier geht es um Hypothesen des Typs $H_0 : \theta \leq \theta_0$ oder $H_0 : \theta \geq \theta_0$ mit einem fixierten θ_0, wir werden nur den ersten Fall behandeln, der andere geht analog.

Wir erinnern uns daran, dass es bei den konkreten Beispielen im vorigen Abschnitt nur auf H_0, nicht aber auf die spezielle Gestalt von H_1 ankam. Das wollen wir uns jetzt zunutze machen:

Satz 3.3.1. *Gegeben sei eine Potenzreihenfamilie, wir verwenden die Bezeichnungen aus Definition 2.3.7 und Satz 2.3.8. Die Nullhypothese habe die Form $H_0 : \theta \leq \theta_0$, weiter sei α eine vorgegebene Fehlerwahrscheinlichkeit. Dann gibt es einen Neyman-Pearson-Test d zum Niveau α, der gleichmäßig in $\theta > \theta_0$ die kleinsten Wahrscheinlichkeiten für Fehler zweiter Art realisiert.*

Er hat die folgende Form: Man kann Zahlen $c \in \mathbb{R}$ und $\delta \in [\,0, 1\,]$ so wählen, dass d auf den Mengen $\{T < c\}$ bzw. $\{T > c\}$ bzw. $\{T = c\}$ den Wert 0 bzw. 1 bzw. δ hat.

Beweis: Wir wählen c und δ so, dass die wie vorstehend definierte Funktion d unter θ_0 den Erwartungswert α hat. Aus dem vorigen Abschnitt wissen wir, dass d ein optimaler Test (= kleinster Fehler zweiter Art) zum Niveau α für θ_0 gegen θ_1 für alle $\theta_1 > \theta_0$ ist.

Nun ein kleiner Trick: Wir betrachten $d' := 1 - d$. Der Erwartungswert ist $1-\alpha$, und es ist ein Neyman-Pearson-Test: Es ist der Neyman-Pearson Test von θ_0 gegen θ', wo $\theta' < \theta_0$ beliebig ist.

Neyman-Pearson-Tests sind bestmöglich. Wenn man also als Variante zu diesem Testproblem die konstante Funktion $d'' := 1 - \alpha$ angibt[6)], so muss die Wahrscheinlichkeit für den Fehler zweiter Art unter d' eher günstiger sein, d.h.:

$$1 - E_{\theta'}(d) = E_{\theta'}(d') \geq E_{\theta'}(d'') = 1 - \alpha.$$

So folgt

$$E_{\theta'}(d) \leq E_{\theta_0}(d)$$

für $\theta' \leq \theta_0$: Der Erwartungswert ist also monoton steigend in θ, und deswegen reicht es statt alle $\theta \leq \theta_0$ nur θ_0 selbst zu untersuchen. Da der Test optimal für θ_0 gegen alle $\theta_1 > \theta_0$ ist, ist die Behauptung bewiesen. □

Es ist nicht erforderlich, neue Beispiele anzugeben, da die des vorigen Abschnitts eigentlich schon Tests für einseitige Hypothesen waren. So ist etwa d im konkreten Beispiel zur Binomialverteilung schon optimaler Test für $p \leq 0.5$ gegen $p > 0.5$.

Zweiseitige Tests

Hier geht es um Hypothesen der Form $\theta = \theta_0$, man möchte sich also gleichzeitig gegen $\theta < \theta_0$ und $\theta > \theta_0$ absichern. Wenn man dann alle Tests zulässt, für die der Fehler erster Art Wahrscheinlichkeit α hat, so kann es darunter in der Regel keinen gleichmäßig besten geben: Man kann ja welche konstruieren, für die der Fehler zweiter Art im Bereich $\theta < \theta_0$ gleichmäßig kleinstmöglich ist – er ist dann für die $\theta > \theta_0$ groß – oder umgekehrt. Eher hat man das Gefühl, dass man einen Kompromiss eingehen müsste.

Um weiterzukommen, führt man eine neue Definition ein: Man beschränkt sich auf Tests, für die der Fehler 2. Art gleichmäßig höchstens $1 - \alpha$ ist. (Das leistet zum Beispiel schon die konstante Abbildung $x \mapsto \alpha$.)

Definition 3.3.2. *Ein Testproblem sei durch die Aufteilung von Θ in die disjunkte Vereinigung von Θ_0 und Θ_1 definiert. Ein randomisierter Test $d : \Omega^n \to [\,0, 1\,]$ heißt* unverfälscht zum Niveau α, *wenn*

$$E_{\theta_0}(d) \leq \alpha \leq E_{\theta_1}(d)$$

für alle $\theta_0 \in \Theta_0$ und $\theta_1 \in \Theta_1$.

[6)] Auch das ist ein Test zum Niveau $1 - \alpha$.

Satz 3.3.3. *Wieder verwenden wir die Bezeichnungen zu Potenzreihenfamilien aus Abschnitt 2.3 (vgl. Definition 2.3.7 und Satz 2.3.8). Zu Testen sei die Nullhypothese $\theta = \theta_0$ gegen die Alternative $\theta \neq \theta_0$, das Niveau α sei vorgegeben. d sei ein Test, es gebe Zahlen c_1, c_2, und dafür gelte:*

(i) $E_{\theta_0}(d) = \alpha$.

(ii) *d hat auf $\{T < c_1\} \cup \{T > c_2\}$ den Wert 1 und auf $\{c_1 < T < c_2\}$ den Wert 0.*

Dann ist d ein gleichmäßig bester unverfälschter Test zum Niveau α.

Beweis: Den findet man im Buch von Irle ab Seite 337. □

Tests im Zusammenhang mit der Normalverteilung

Im Spezialfall normalverteilter Zufallsvariablen lassen sich beste Tests leicht explizit angeben, die Symmetrie der Normalverteilung führt noch zu einer gewissen Vereinfachung. Wir besprechen hier einige ausgewählte Beispiele, die Einzelheiten sind im Buch von Irle in Kapitel 20 ausgeführt.

A. *Einseitiger Gaußtest*

σ_0 sei bekannt, das statistische Modell bestehe aus allen $N(a, \sigma_0^2)$, und wir wollen $a \leq a_0$ für ein vorgegebenes a_0 testen. (Etwa: Sind die Erwartungswerte gewisser normalverteilter Längen durch a_0 beschränkt?)

Das Verfahren: Bestimme bei vorgegebenem α ein r, so dass $[\, r, +\infty [$ unter $N(0,1)$ die Wahrscheinlichkeit α hat. Lehne dann aufgrund der Messungen $x_1, \ldots, x_n$ die Hypothese $H_0 : a \leq a_0$ genau dann ab, wenn

$$\sqrt{n}\, \frac{\overline{x} - a_0}{\sigma_0} > r.$$

B. *Zweiseitiger Gaußtest*

σ_0 sei wieder bekannt, das statistische Modell bestehe aus allen $N(a, \sigma_0^2)$. Diesmal wollen wir die Nullhypothese $a = a_0$ gegen $a \neq a_0$ testen.

Das Verfahren: Bestimme bei vorgegebenem α ein r, so dass $[\, -r, +r\,]$ unter $N(0,1)$ die Wahrscheinlichkeit $1 - \alpha$ hat. Lehne dann aufgrund der Messungen $x_1, \ldots, x_n$ die Hypothese $H_0 : a \neq a_0$ genau dann ab, wenn

$$\left| \sqrt{n}\, \frac{\overline{x} - a_0}{\sigma_0} \right| > r.$$

C. *Einseitiger t-Test*

Diesmal sind *alle* $N(a, \sigma^2)$ zugelassen, es soll $a \leq a_0$ für ein vorgegebenes a_0 getestet werden.

Das Verfahren: Bestimme bei vorgegebenem α ein r, so dass $[\, r, +\infty\, [$ unter t_{n-1} die Wahrscheinlichkeit α hat. Lehne dann aufgrund der Messungen $x_1, \ldots, x_n$ die Hypothese $H_0 : a \leq a_0$ genau dann ab, wenn

$$\sqrt{n}\ \frac{\overline{x} - a_0}{\sqrt{V_x}} > r.$$

D. *Zweiseitiger t-Test*

Wie vorstehend, aber die Nullhypothese lautet $a = a_0$.

Das Verfahren: Bestimme bei vorgegebenem α ein r, so dass $[\, -r, +r\,]$ unter t_{n-1} die Wahrscheinlichkeit $1 - \alpha$ hat. Lehne dann aufgrund der Messungen $x_1, \ldots, x_n$ die Hypothese $H_0 : a = a_0$ genau dann ab, wenn

$$\left| \sqrt{n}\ \frac{\overline{x} - a_0}{\sqrt{V_x}} \right| > r.$$

E. *Einseitiger χ^2-Test*

Zur Konkurrenz sind alle $N(a_0, \sigma^2)$ bei unbekanntem a_0 zugelassen, und wir wollen $\sigma \leq \sigma_0$ testen.

Das Verfahren: Bestimme bei vorgegebenem α ein r, so dass $[\, r, +\infty\, [$ unter χ^2_{n-1} die Wahrscheinlichkeit α hat. Lehne dann aufgrund der Messungen $x_1, \ldots, x_n$ die Hypothese $H_0 : \sigma \leq \sigma_0$ genau dann ab, wenn

$$(n-1)\frac{V_x}{\sigma_0^2} > r.$$

F. *Zweiseitiger χ^2-Test*

Wie vorstehend, es soll die Nullhypothese $\sigma = \sigma_0$ gegen $\sigma \neq \sigma_0$ getestet werden.

Das Verfahren: Bestimme bei vorgegebenem α Zahlen $0 < r_1 < r_2$, so dass $[\, r_1, r_2\,]$ unter χ^2_{n-1} die Wahrscheinlichkeit $1 - \alpha$ hat und die r_1, r_2 zu den Bedingungen von Satz 3.3.3 passen. Das kann kompliziert sein, im Zweifelsfall sollte man sich auf statistische Software verlassen.
Lehne dann aufgrund der Messungen $x_1, \ldots, x_n$ die Hypothese $H_0 : \sigma = \sigma_0$ genau dann ab, wenn

$$(n-1)\frac{V_x}{\sigma_0^2} \notin [\, r_1, r_2\,].$$

Kapitel 4

Lineare Modelle

In diesem Kapitel geht es um Situationen, in denen die beobachteten Größen von der Form

„Lineare Funktion der Eingabedaten plus zufällige Störung“

sind. Das ist ein relativ spezieller Ansatz, der aber überraschend viele Anwendungen hat. Wir beginnen in *Abschnitt 4.1* mit einer Motivation, die uns zur allgemeinen Form linearer Modelle führt. In diesem allgemeinen Rahmen untersuchen wir, wie man zu vernünftigen optimalen Schätzern kommen kann, das Hauptergebnis wird der *Satz von Gauß-Markov* sein. Bemerkenswerterweise haben alle hier relevanten Ergebnisse eine einfache *geometrische Interpretation*. Das liegt daran, dass die Probleme bei geschickter Übersetzung auf Fragen in euklidischen Räumen führen, eine wichtige Rolle werden deswegen auch Methoden aus der Linearen Algebra spielen.

Die Theorie wird in Abschnitt 4.1 für den Fall entwickelt, dass eine gewisse für das Modell wichtige Matrix invertierbar ist.Später wird aber auch eine allgemeinere Situation wichtig werden. Um auch für diesen Fall Ergebnisse erhalten zu können, wird die Theorie in *Abschnitt 4.2* erweitert, vorbereitend wird eine Konstruktion aus der Linearen Algebra behandelt, die Pseudoinverse.

Dann gibt es einen wahrscheinlichkeitstheoretischen Exkurs: In *Abschnitt 4.3* kümmern wir uns um *mehrdimensionale Normalverteilungen*. Die haben überraschende Eigenschaften, insbesondere wird wichtig sein, dass man – wie in Abschnitt 2.5 – sehr detaillierte Aussagen über die daraus abgeleiteten Verteilungen machen kann. In *Abschnitt 4.4* werden dann die allgemeinen Ergebnisse über lineare Modelle für den Spezialfall untersucht, in dem die Störungen gemeinsam normalverteilt sind.

Als Nächstes gehen wir in *Abschnitt 4.5* noch einmal genauer auf das Thema „Schätzen“ ein. Insbesondere soll gezeigt werden, dass die unter der Normalverteilungsannahme konstruierten Schätzer in einem präzisen Sinn bestmöglich sind.

In *den letzten zwei Abschnitten* werden dann, zum Abschluss des Kapitels, zwei wichtige Anwendungsklassen linearer Modelle behandelt, nämlich *Varianzanalyse* und *Kovarianzanalyse*.

4.1 Das lineare Modell 1 (voller Rang der Designmatrix)

Die einfachsten Abbildungen nach den konstanten sind die linearen Abbildungen. Sie spielen deswegen eine wichtige Rolle, weil sie einerseits noch gut zu beherrschen sind, andererseits aber auch häufig in den Anwendungen auftreten. So hat zum Beispiel das Konzept „Differenzierbarkeit“ die Konsequenz, dass differenzierbare Abbildungen im Kleinen wie lineare Abbildungen behandelt werden können. Auch physikalische Zusammenhänge sind oft näherungsweise linear, man denke nur an das Hookesche Gesetz.

Allerdings kann man lineare Zusammenhänge durch konkrete Messungen nie exakt feststellen, da es immer Fehlereinflüsse gibt. Die liegen zum Teil im Messprozess, zum Teil aber auch daran, dass ein eigentlich linearer Vorgang durch unbekannte Störungen verändert wird.

4.1.1 Die allgemeine Definition

Zunächst sind einige *Vokabeln* zu besprechen. Es gebe gewisse Eingangsgrößen $\gamma_1, \ldots, \gamma_s$, die durch einen linearen Vorgang Ausgangsgrößen $X_1, \ldots, X_n$ erzeugen[1)]. Jedenfalls im Wesentlichen: Genau genommen wird die lineare Operation noch durch eine Störung überlagert.

Gegeben ist zunächst eine reelle $n \times s$-Matrix A, die so genannte *Designmatrix*. In diesem Abschnitt soll angenommen werden, dass sie vollen Rang hat, dadurch werden viele Überlegungen wesentlich einfacher. Der Zufall wird so modelliert: Es gibt eine Familie $\xi_1, \ldots, \xi_n$ (die *Störgrößen*) von unabhängigen, identisch verteilten Zufallsvariablen, die alle Erwartungswert Null und Streuung Eins haben, und die Verfälschung der k-ten X-Messung wird durch einen Summanden der Form $\sigma\xi_k$ ausgedrückt, wobei σ die in der Regel unbekannte Streuung der Zufallsvariablen $\sigma\xi_k$ ist.

Führt man noch Spaltenvektoren $X := (X_1, \ldots, X_n)^\top$, $\gamma := (\gamma_1, \ldots, \gamma_s)^\top$ und $\xi := (\xi_1, \ldots, \xi_n)^\top$ ein, so kann man alles kompakt in der Formel

$$X = A\gamma + \sigma\xi$$

zusammenfassen. *Das* ist das lineare Modell.

Dabei sind $\gamma \in \mathbb{R}^s$ und meist auch $\sigma \in \mathbb{R}$ unbekannt. In der Sprechweise der statistischen Modelle aus Abschnitt 2.1 haben wir also eine durch γ und σ parametrisierte Schar von Situationen vor uns, und die Aufgabe besteht darin, durch konkrete X-Realisierungen Aussagen über γ und σ zu gewinnen.

[1)] Die γ_i heißen manchmal die *Regressoren*, die X_i die *Regressanden*. Wir wollen annehmen, dass $s \leq n$ gilt.

Man kann das Problem auch geometrisch deuten. Dazu bezeichnen wir die s Spaltenvektoren von A mit $z_1, \ldots, z_s$. Wenn A vollen Rang hat, sind sie linear unabhängig. Dann ist $A\gamma$ eine – durch eine Störung etwas verfälschte – Linearkombination dieser Vektoren, und die Aufgabe besteht darin, die zugehörigen Koeffizienten und die Stärke der Störung möglichst genau zu identifizieren.

Es ist klar, dass der Schwierigkeitsgrad von der gegenseitigen Lage der z_i abhängen wird: Wenn sie senkrecht aufeinander stehen, sind sicher bessere Ergebnisse zu erwarten, als wenn die Winkel zwischen ihnen sehr klein sind.

Mit Hilfe der nun zu entwickelnde Theorie werden wir in der Lage sein, eine auf präzisierbare Weise optimale quantitative Lösung zu finden.

Beispiele

1. Angenommen, es liegt eine $N(\mu, \sigma^2)$ verteilte Zufallsvariable vor, über die durch n Messungen Informationen gewonnen werden sollen. Für die k-te Messung gilt also die Gleichung

$$X_k = \mu + \sigma\xi_k,$$

wobei ξ_k $N(0,1)$-verteilt ist. Wählt man die Designmatrix A als $(1, \ldots, 1)^\top$ und $\gamma = (\mu)$, so erweist sich diese Situation als Spezialfall eines linearen Modells (wobei hier $s = 1$ ist).

2. Die Suche nach einer Regressionsgeraden aus Abschnitt 1.5 kann nun auch noch einmal aufgegriffen werden. Zur Erinnerung: Es waren Tupel $(x_k, y_k) \in \mathbb{R}^2$ für $k = 1, \ldots, n$ (eine „Punktwolke“) gegeben, und gesucht war eine Gerade $x \mapsto \gamma_0 + \gamma_1 x$, die diese Punkte möglichst gut interpoliert.

Wir können das so interpretieren: „Eigentlich“ ist der zu x_k gehörige Wert gleich $\gamma_0 + \gamma_1 x_k$, aber auf Grund unvorhergesehener Störungen wird der zu $\gamma_0 + \gamma_1 x + \sigma\xi_k$ verändert.

Das ist ein Spezialfall des linearen Modells, wenn wir als Designmatrix die $n \times 2$-Matrix A wählen, die in der k-ten Zeile die Werte $(1, x_k)$ hat. γ ist hier der Vektor $(\gamma_0, \gamma_1)^\top$, und die Messungen lassen sich als Matrixgleichung

$$(y_1, \ldots, y_n)^\top = A\gamma + \sigma\xi$$

zusammenfassen.

Das wollen wir nun *geometrisch interpretieren*. A entspricht doch einer Abbildung vom $\mathbb{R}^2$ in den $\mathbb{R}^n$. Wir sind an γ interessiert, doch da A injektiv ist, ist dazu die Kenntnis von $A\gamma$ gleichwertig. Jede konkrete Messung liefert nur $A\gamma + \sigma\xi$, und man kann sich fragen, wie man da $A\gamma$ optimal schätzen sollte.

Nun liegt $A\gamma$ in dem zweidimensionalen Unterraum $A(\mathbb{R}^2)$, und man kann überlegen, welcher Punkt dieses Unterraums aufgrund der konkreten Realisierung $A\gamma+\sigma\xi$ ausgewählt werden sollte. Die Antwort: Derjenige, der am nächsten dran ist! Man sollte also als Schätzung für γ dasjenige $\hat{\gamma} = (\hat{\gamma_0}, \hat{\gamma_1})$ wählen, für das $\left\|A\hat{\gamma} - (y_1, \ldots, y_n)^\top\right\|$ so klein wie möglich ist.

Satz 4.1.1. *Wir verwenden die vorstehenden Bezeichnungen. Setze*

$$\begin{aligned}
\overline{x} &:= (x_1 + \cdots + x_n)/n, \\
\overline{y} &:= (y_1 + \cdots + y_n)/n, \\
c &:= (x_1 y_1 + \cdots + x_n y_n)/n - \overline{x}\,\overline{y}, \\
v_x &:= \sum_k (x_k - \overline{x})^2/n, \\
\hat{\gamma_0} &:= \overline{y} - \frac{\overline{x}}{v_x} c, \\
\hat{\gamma_1} &:= \frac{c}{v_x}.
\end{aligned}$$

(Beachte: v_x unterscheidet sich nur unwesentlich von der Stichprobenvarianz V_x aus Abschnitt 2.1, da stand $n-1$ im Nenner.) Dann gilt:

(i) $x \mapsto \hat{\gamma_0} + \hat{\gamma_1} x$ ist die Regressionsgerade, die wir nach der Formel aus Abschnitt 1.5 zu den Tupeln (x_k, y_k) ausgerechnet hätten.

(ii) $(\hat{\gamma_0}, \hat{\gamma_1})$ ist auch derjenige Schätzwert für (γ_0, γ_1), der sich aufgrund der vorstehend beschriebenen Strategie ergeben würde.

(iii) Beide Schätzungen ($\hat{\gamma_0}$ für γ_0 und $\hat{\gamma_1}$ für γ_1) sind erwartungstreu.

Beweis: (i) In Satz 1.3.5 hatten wir bewiesen: Ist $\overline{x} = \overline{y} = 0$, so ist diejenige Gerade $a + bx$, für die sich die kleinste Summe der quadrierten Abstände ergibt, durch $a = 0$ und $b = \sum x_k y_k / \sum x_k^2$ gegeben.

Um dieses Ergebnis mit der hier betrachteten allgemeineren Situation zu vergleichen, setzen wir $x'_k := x_k - \overline{x}$ und $y'_k := y_k - \overline{y}$ und wenden Satz 1.3.5 auf die x'_k, y'_k an. Die beste Näherung für die y'_k durch eine Gerade der Form $a + bx'_k$ ist also durch bx'_k mit $b = \sum x'_i y'_i / \sum (x'_i)^2$ gegeben.

Nun schreiben wir bx'_kin der Form $\alpha + \beta x_k$:

$$\begin{aligned}
y'_k &= y_k - \overline{y} \\
&\approx bx'_k \\
&= \frac{\sum (x_i - \overline{x})(y_i - \overline{x})}{\sum (x_i - \overline{x})^2} (x_k - \overline{x}) \\
&= -\frac{c}{v_x}\overline{x} + \frac{c}{v_x} x_k,
\end{aligned}$$

also

$$y_k \approx \hat{\gamma_0} + \hat{\gamma_1} x_k.$$

(ii) Das ist klar, denn „Die Summe der Abweichungsquadrate für die approximierende Gerade $\hat{\gamma_0} + \hat{\gamma_1} x$ ist minimal“ ist nach Übersetzung in die Geometrie des $\mathbb{R}^n$ die Aussage „Der Punkt $A\gamma$ hat unter allen Punkten im Bild von A den kleinsten Abstand zu $(y_1, \ldots, y_n)$“.

(iii) Es ist zu zeigen: Wenn die *wirklichen* Parameter γ_0 und γ_1 sind und der Zufall zu den Werten $y_k = \gamma_0 + \gamma_1 x_k + \sigma\xi_k$ führt, so wird das obige Schätzverfahren ($\hat{\gamma_0}$ für γ_0 und $\hat{\gamma_1}$ für γ_1) im Mittel zu den richtigen γ_0, γ_1 führen.

Da der Erwartungswert der ξ_k gleich Null ist, folgt $E(y_k) = \gamma_0 + \gamma_1 x_k$, und damit ist auch $E(\overline{y}) = \gamma_0 + \gamma_1 \overline{x}$. Für die Berechnung des Erwartungswertes von $\hat{\gamma_1}$ nutzen wir die Linearität aus:

$$\begin{aligned} E(\hat{\gamma_1}) &= \frac{1}{v_x}\big(\frac{1}{n}\sum_k x_k(\gamma_0 + \gamma_1 x_k) - \overline{x}(\gamma_0 + \gamma_1 \overline{x})\big) \\ &= \frac{1}{v_x}\left(\gamma_0\overline{x} + \gamma_1 \frac{1}{n}\sum_k \big(x_k^2 - \overline{x}^2\big) - \overline{x}\gamma_0\right) \\ &= \frac{1}{v_x}\big(\gamma_1 v_x\big) \\ &= \gamma_1. \end{aligned}$$

Und weiter:

$$\begin{aligned} E(\hat{\gamma_0}) &= \gamma_0 + \gamma_1\overline{x} - \frac{\overline{x}}{v_x}E(c) \\ &= \gamma_0 + \gamma_1\overline{x} - \frac{\overline{x}}{v_x}v_x\gamma_1 \\ &= \gamma_0. \end{aligned}$$

□

4.1.2 Designmatrix mit vollem Rang: Schätzen der Parameter

Der Ansatz, der sich gerade im Fall der Regression als erfolgreich erwiesen hat, soll nun auf die allgemeine Situation übertragen werden. Wir beginnen mit einem Ergebnis aus der Theorie der euklidischen Räume:

Lemma 4.1.2. *Es sei U ein Unterraum des $\mathbb{R}^n$, der mit dem üblichen Skalarprodukt versehen sei. Für ein $x_0 \in \mathbb{R}^n$ und ein $y_0 \in U$ sind äquivalent:*

(i) $\|x_0 - y_0\| = \min_{y\in U} \|x_0 - y\|$.

(ii) $x_0 - y_0$ *steht senkrecht auf* U.

y_0 ist durch diese Eigenschaft eindeutig bestimmt, wir nennen es die Projektion von x_0 auf U *und schreiben $P_U x_0 := y_0$.*

Beweis: y_0 möge den Minimalabstand realisieren, o.B.d.A. sei $y_0 = 0$. Für $y \in U$ ist folglich $\|x_0 - y\| \geq \|x_0\|$. Andererseits ist nach dem Satz von Pythagoras

$$\|a + b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b\rangle.$$

In unserem Fall gilt also, wenn wir $y \in U$ vorgeben,

$$\|x_0\|^2 \leq \|x_0 - ty\|^2 = \|x_0\|^2 + t^2\|y\|^2 - 2t\langle x_0, y\rangle$$

für jedes t. Aber $t^2\alpha - 2t\beta$ ist nur dann für alle t positiv, wenn $\beta = 0$ gilt. Also ist $\langle x_0, y\rangle = 0$.

Umgekehrt: x_0 stehe senkrecht auf allen $y \in U$. Wir zeigen, dass bei 0 der kleinstmögliche Abstand realisiert wird: Nach dem Satz von Pythagoras gilt nämlich $\|x_0 - y\|^2 = \|x_0\|^2 + \|y\|^2 \geq \|x_0\|^2$.

Es fehlt noch der Nachweis der Eindeutigkeit[2)]. Sind y_0 und y_0' Elemente bester Approximation, so gilt $\|x_0 - y_0\| = \|x_0 - y_0'\|$. Es steht aber $y_0 - y_0'$ senkrecht auf $x_0 - y_0'$, nach dem Satz von Pythagoras ist also

$$\|x_0 - y_0\|^2 = \|(x_0 - y_0') + (y_0' - y_0)\|^2 = \|x_0 - y_0'\|^2 + \|y_0' - y_0\|^2.$$

Folglich ist $\|y_0' - y_0\|^2 = 0$, d.h. $y_0 = y_0'$. □

Uns interessiert besonders der Spezialfall, in dem U der Bildraum einer linearen Abbildung ist. Da lässt sich P_U explizit durch eine Matrix angeben:

Satz 4.1.3. *Es sei A eine $n \times s$-Matrix, die zugehörige lineare Abbildung vom $\mathbb{R}^s$ in den $\mathbb{R}^n$ bezeichnen wir ebenfalls mit A. Es soll $s \leq n$ sein, und wir setzen voraus, dass A Rang s hat. Bezeichnet man mit $A^\top$ die zu A transponierte Matrix, so gilt:*

(i) $A^\top A$ ist invertierbar.

(ii) Sei $U := \{A\gamma \mid \gamma \in \mathbb{R}^s\}$ der Bildraum von A. Dann entspricht die Projektion P_U auf U der folgenden Matrix:

$$\Pi_A := A\big(A^\top A\big)^{-1}A^\top.$$

Es ist also $P_U x = \Pi_A x$ für jedes x, und insbesondere ist P_U linear.

Beweis: Wir werden häufig von der folgenden Identität Gebrauch machen, die sich unmittelbar aus der Definition ergibt:

$$\langle A\gamma, x\rangle = \langle \gamma, A^\top x\rangle.$$

(i) Sei $\gamma \neq 0$. Dann ist, da A nach Voraussetzung injektiv ist, auch $A\gamma \neq 0$. Es folgt

$$0 \neq \|A\gamma\|^2 = \langle A\gamma, A\gamma\rangle = \langle A^\top A\gamma, \gamma\rangle,$$

also ist $A^\top A\gamma \neq 0$. Als injektive Abbildung auf einem endlich-dimensionalen Raum muss $A^\top A$ damit auch bijektiv sein.

(ii) Sei x_0 beliebig. Wir definieren y_0 durch $A\big(A^\top A\big)^{-1}A^\top x_0$ und müssen zeigen, dass y_0 das Element bester Approximation ist.

[2)] Genau genommen fehlt auch noch der Nachweis der Existenz. Das geht am einfachsten mit einem Kompaktheitsargument: U ist abgeschlossen, $y \mapsto \|y - x_0\|$ ist stetig auf U, und es reicht, das Minimum auf einer genügend großen Kugel – die dann kompakt ist – zu suchen.

Klar ist, dass y_0 in U liegt, denn $y_0 = A\gamma_0$, wo $\gamma_0 = \big(A^\top A\big)^{-1} A^\top x_0$. Aufgrund des vorigen Lemmas ist zu zeigen, dass $x_0 - y_0$ auf allen Elementen von U senkrecht steht. Zum Beweis geben wir ein beliebiges $A\gamma \in U$ vor und schließen so:

$$\begin{aligned}\langle A\gamma, x_0\rangle &= \langle \gamma, A^\top x_0\rangle \\ &= \langle \gamma, \big(A^\top A\big)\big(A^\top A\big)^{-1} A^\top x_0\rangle \\ &= \langle A\gamma, A\big(A^\top A\big)^{-1} A^\top x_0\rangle \\ &= \langle A\gamma, y_0\rangle.\end{aligned}$$

Das bedeutet wirklich $\langle A\gamma, x_0 - y_0\rangle = 0$, und damit ist y_0 *das* Element bester Approximation. □

Wir wenden uns nun wieder dem allgemeinen linearen Modell $X = A\gamma + \sigma\xi$ zu. Wie kann man optimal das γ schätzen, wenn X gemessen wurde? Aufgrund der Überlegungen im Zusammenhang mit der Regression sollte die Schätzung $\hat\gamma$ von γ derjenige Vektor $\hat\gamma$ sein, für den $A\hat\gamma$ kleinstmöglichen Abstand zu X hat. Wirklich gilt:

Satz 4.1.4. *(Satz von Gauß-Markov, Designmatrix mit vollem Rang, Teil 1) Im linearen Modell werden die besten Schätzungen so beschrieben:*

(i) Derjenige Vektor $\hat\gamma$, für den $A\hat\gamma$ kleinstmöglichen Abstand zu X hat, ist durch $\hat\gamma := \big(A^\top A\big)^{-1} A^\top X$ gegeben[3)]*.*

Folglich ist $\hat\gamma$ dadurch charakterisiert, dass $X - A\hat\gamma$ senkrecht auf dem Bild von A steht, für alle γ also die Gleichung

$$\langle X - A\hat\gamma, A\gamma\rangle = 0$$

gilt. Diese Gleichungen heißen die Normalgleichungen*.*

(ii) $X \mapsto \hat\gamma$ ist ein erwartungstreuer linearer Schätzer[4)] *für γ.*

(iii) Die Kovarianzmatrix dieses Schätzers ist $\sigma^2 (A^\top A)^{-1}$.

(iv) Die Streuung von $\hat\gamma$, also der Erwartungswert von $||\gamma - \hat\gamma||^2$ (dem Abstandsquadrats zwischen wahrem Wert und Schätzung), ist gleich $\sigma^2 Spur\,(A^\top A)^{-1}$.

(v) Der Schätzer $\hat\gamma$ ist in folgendem Sinn optimal: Ist B eine $s \times n$-Matrix und $X \mapsto BX$ ein weiterer linearer erwartungstreuer Schätzer, so ist die Streuung von B (der Erwartungswert von $||\gamma - B\big(A\gamma + \xi\big)||^2$ nicht kleiner als $\sigma^2 Spur\,(A^\top A)^{-1}$.

Und Gleichheit gibt es nur im Fall $B = (A^\top A)^{-1} A^\top$

[3)] Man mache sich klar, dass dieses Vorgehen einer höherdimensionalen Verallgemeinerung der Technik entspricht, die zur Regressionsgeraden geführt hat: Die Quadratsumme der Abweichungen bei einer gewissen Abbildung einer Teilmenge des $\mathbb{R}^s$ nach $\mathbb{R}$ soll in der Klasse der linearen Abbildungen minimiert werden.

[4)] Genauer: Der Schätzer selbst ist die Abbildung $x \mapsto \big(A^\top A\big)^{-1} A^\top x$ von $\mathbb{R}^n$ nach $\mathbb{R}^s$. Sie wird auf Zufallsvektoren angewendet.

(vi) Bezeichnet U das Bild von A, so gilt: Der Vektor $P_U X$ steht senkrecht auf $X - P_U X$, und deswegen ist

$$\|X\|^2 = \|X - P_U X + P_U X\|^2 = \|X - P_U X\|^2 + \|P_U X\|^2.$$

Wir behaupten, dass

$$V^* := \frac{\|X\|^2 - \|P_U X\|^2}{n-s} = \frac{\|X - P_U X\|^2}{n-s}$$

ein erwartungstreuer Schätzer für σ^2 ist.

Beweis: (i) Aufgrund des vorstehenden Satzes wissen wir, dass $A\big(A^\top A\big)^{-1} A^\top X$ das Element bester Approximation an X ist. Diesen Vektor kann man aber als $A\hat{\gamma}$ schreiben.

(ii) Die Linearität ist klar, für den Beweis der Erwartungstreue nutzen wir die Linearität des Erwartungswertes aus:

$$\begin{aligned} E(\hat{\gamma}) &= E\Big(\big(A^\top A\big)^{-1} A^\top X\Big) \\ &= E\Big(\big(A^\top A\big)^{-1} A^\top (A\gamma + \sigma\xi)\Big) \\ &= \Big(\big(A^\top A\big)^{-1} A^\top (A\gamma)\Big) + \sigma E\Big(\big(A^\top A\big)^{-1} A^\top (\xi)\Big) \\ &= \gamma + \sigma\big(\big(A^\top A\big)^{-1} A^\top E(\xi) \\ &= \gamma. \end{aligned}$$

Es wurde nur gebraucht, dass der Erwartungswert von ξ (komponentenweise) gleich Null ist.

(iii) Sei $D = (d_{i,j})$ eine $s \times n$-Matrix. Die i-te bzw. k-te Komponente des Zufallsvektors $D\xi$ sind dann durch $\sum_j d_{ij}\xi_j$ bzw. $\sum_l d_{kl}\xi_l$ gegeben. Damit ist der Eintrag an der Stelle (i,k) der Kovarianzmatrix der Erwartungswert von

$$\sum_{j,l} d_{ij}\xi_i d_{kl}\xi_l,$$

also gleich

$$\sum_j d_{ij} d_{kj};$$

hier wurde ausgenutzt, dass die Komponenten von ξ unkorrelliert und identisch verteilt sind. Diese Zahl ist aber der Eintrag bei (i,k) in $DD^\top$.

Hier ist das für $D = \left(A^\top A\right)^{-1} A^\top$ anzuwenden; dabei wird es gleich wichtig werden, dass $A^\top A$ und folglich auch $(A^\top A)^{-1}$ eine symmetrische Matrix ist:

$$\begin{aligned}\sigma^2 DD^\top &= \sigma^2 (A^\top A)^{-1} A^\top \left((A^\top A)^{-1} A^\top\right)^\top \\ &= \sigma^2 (A^\top A)^{-1} A^\top A \left((A^\top A)^{-1}\right)^\top \\ &= \sigma^2 (A^\top A)^{-1} A^\top A (A^\top A)^{-1} \\ &= \sigma^2 (A^\top A)^{-1}.\end{aligned}$$

(iv) Wieder beginnen wir mit einer allgemeinen Überlegung. Es ist eine $s \times n$-Matrix D gegeben, diesmal sind wir am Erwartungswert von $||D\xi||^2$ interessiert:

$$\begin{aligned}||D\xi||^2 &= \langle D\xi, D\xi\rangle \\ &= \langle \xi, D^\top D\xi\rangle.\end{aligned}$$

Aus den Bedingungen an ξ folgt nun sofort, dass für quadratische Matrizen C der Erwartungswert von $\langle \xi, C\xi\rangle$ gleich dem σ^2-fachen der Spur von C ist.

Bei uns ist $D = (A^\top A)^{-1} A^\top$, es ist also die Spur von

$$D^\top D = \left((A^\top A)^{-1} A^\top\right)^\top (A^\top A)^{-1} A^\top$$

von Interesse. Das sieht recht kompliziert aus, auch wenn man diesen Ausdruck unter Ausnutzung der Selbstadjungiertheit von $(A^\top A)^{-1}$ zu

$$A(A^\top A)^{-1}(A^\top A)^{-1} A^\top$$

umformt. Jetzt hilft ein Ergebnis über Spuren weiter:

> Ist C eine $n \times s$-Matrix und D eine $s \times n$-Matrix, so sind die Spuren von CD und DC gleich[5)].

Wir wenden es mit $C = A$ und $D = (A^\top A)^{-1}(A^\top A)^{-1} A^\top$ an. Die gesuchte Spur ist also gleich der Spur von $(A^\top A)^{-1}$. Das beweist die Behauptung, wenn man noch den Faktor σ^2 berücksichtigt.

(v) Dieser Beweis wird vorläufig vertagt. Wir werden das Ergebnis später als Korollar zu einer Aussage über das Schätzen von Funktionen erhalten (s.u., Korollar 4.1.6; vgl. auch Satz 4.2.7 (iii)).

(vi) Die Beweisstrategie besteht darin, die Aussage auf den Fall zurückzuführen, dass der s-dimensionale Unterraum U von den ersten s Einheitsvektoren aufgespannt wird.

Wir beginnen damit, dass wir eine Orthonormalbasis $u_1, \ldots, u_n$ im $\mathbb{R}^n$ so bestimmen, dass die $u_1, \ldots, u_s$ den Raum U aufspannen. Wir fassen die u_k als Spalten einer Matrix $O = (O_{ij})$ auf, O ist dann eine orthogonale Matrix. Mit

$$W := \{(x_1, \ldots, x_s, 0, \ldots, 0) \mid x_1, \ldots, x_s \in \mathbb{R}\}$$

[5)] Das kann man leicht nachrechnen: Beide Spuren sind gleich $\sum_{i=1,\ldots,n, j=1,\ldots,s} c_{ij} d_{ji}$.

ist W ein s-dimensionaler Unterraum des $\mathbb{R}^n$, und $x \mapsto Ox$ vermittelt eine Bijektion auf $\mathbb{R}^n$, die W auf U abbildet. Die orthogonale Projektion von $\mathbb{R}^n$ auf W wird durch diejenige Diagonalmatrix D_s beschrieben, die an den ersten s Diagonaleinträgen den Wert 1 und sonst lauter Nullen hat.

Wir betrachten nun die Matrix $D' := OD_sO^\top$. Sie ist zu D_s konjugiert und damit ebenfalls eine orthogonale Projektion. Ihr Bildraum ist U, und deswegen muss $D' = \Pi_A$ gelten: Die sich daraus ergebende Beziehung $O^\top \Pi_A O = D_s$ wird gleich wichtig werden.
Sei nun $\eta := O^\top \xi$ und

$$\eta_k = \sum_j O_{jk}\xi_j$$

die k-te Komponente von η. (Beachte: η_k ist eine Zufallsvariable, η ein Zufallsvektor.)
Nun ist die Länge eines Vektors x gleich der Länge von Ox, denn

$$\|Ox\|^2 = \langle Ox, Ox\rangle = \langle O^\top Ox, x\rangle = \langle x, x\rangle = \|x\|^2.$$

Wir wollen das auf V^* anwenden, dazu müssen wir für $X = A\gamma + \sigma\xi$ den Vektor $X - P_U X$ betrachten. Nach Definition ist

$$\begin{aligned} P_U X &= A\big(A^\top A\big)^{-1}A^\top(A\gamma + \sigma\xi) \\ &= A\gamma + \sigma A\big(A^\top A\big)^{-1}A^\top\xi \\ &= A\gamma + \sigma\Pi_A\xi, \end{aligned}$$

d.h. $X - P_U X = \sigma(\xi - \Pi_A\xi)$.
Insbesondere ist also

$$\begin{aligned} (n-s)V^* &= \|X - P_U X\|^2 \\ &= \sigma^2\|\xi - \Pi_A\xi\|^2 \\ &= \sigma^2\big\|O^\top(\xi - \Pi_A\xi)\big\|^2 \\ &= \sigma^2\big\|\eta - O^\top\Pi_A\xi\big\|^2 \\ &= \sigma^2\|\eta - D_s\eta\|^2 \\ &= \sigma^2\sum_{k=s+1}^{n}\eta_k^2. \end{aligned}$$

Daraus folgt: Der Erwartungswert von $(n-s)V^*$ ist gleich $\sigma^2\sum_{k=s+1}^n E(\eta_k^2)$. Es ist aber

$$\eta_k^2 = \sum_{1\le i,j\le n} O_{ik}O_{jk}\xi_i\xi_j,$$

aus Linearitätsgründen (und weil O orthonormal ist und weil ξ unabhängige und normalisierte Komponenten hat) ist der Erwartungswert also 1. Es folgt

$$E\big((n-s)V^*\big) = (n-s)\sigma^2,$$

und das ist gerade die Behauptung. □

Als erstes *sehr einfaches Beispiel* sollen zur Illustration zwei Situationen verglichen werden:

a) Wir betrachten das Modell $X_1 = \gamma_1 + \sigma\xi_1,\ X_2 = \sigma\xi_2$. Hier ist die Designmatrix $A = (1\ 0)^\top$, und es folgt, dass $X \mapsto X_1$ ein optimaler Schätzer für $\gamma = (\gamma_1)$ ist; die Varianz ist σ^2. Unsere erwartungstreue Schätzung für σ^2 ist X_2^2.

b) Diesmal untersuchen wir das Modell $X_1 = \gamma_1 + \sigma\xi_1,\ X_2 = \gamma_1 + \sigma\xi_2$, es ist also $A = (1\ 1)^\top$. Der beste Schätzer ist $\overline{x} = (X_1 + X_2)/2$, die Varianz ist nur noch $\sigma^2/2$. Und σ^2 wird durch $(X_1 - \overline{x})^2 + (X_2 - \overline{x})^2$ (also die Stichprobenvarianz) geschätzt.

Man beachte, dass wir im zweiten Fall γ_1 besser schätzen konnten; beim ersten Beispiel wurde Information über γ_1 quasi verschenkt, weil diese Zahl in der zweiten Gelichung nicht vorkam.

Durch die eben bewiesenen Ergebnisse haben wir einen Schätzer für γ mit gewissen Optimalitätseigenschaften erhalten. Es gibt aber auch viele Situationen, bei denen gar nicht alle Komponenten von γ interessant sind, sondern nur eine gewisse, sich daraus ergebende Zahl:

- Es könnte zum Beispiel sein, dass man nur Informationen über γ_3 haben möchte.

- Falls in einer Anwendung die γ_i als Einflussgrößen oder Erträge interpretiert werden können, kann die Differenz von – zum Beispiel – γ_1 und γ_2 eine Rolle spielen: Je nachdem, ob sie > 0, $= 0$ oder < 0 ist, hat γ_1 stärkeren, gleichen oder schwächeren Einfluß im Vergleich zu γ_2.

Unsere Theorie wird auf *lineare* Funktionen von γ anwendbar sein. Gegeben ist also eine lineare Abbildung $\Gamma : \mathbb{R}^s \to \mathbb{R}$, und das Problem besteht darin, aufgrund der Messung möglichst gute Schätzungen für $\Gamma(\gamma)$ zu finden. Bevor es weiter geht, sollte man sich daran erinnern, dass lineare Abbildungen durch das Skalarprodukt dargestellt werden können. Zu Γ gibt es also ein eindeutig bestimmtes $z \in \mathbb{R}^s$, so dass $\Gamma\gamma = \langle z, \gamma\rangle$ für alle γ gilt.

Satz 4.1.5. *(Satz von Gauß-Markov, Designmatrix mit vollem Rang, Teil 2) Es sei* Γ *wie vorstehend.*

(i) $T_\Gamma : X \mapsto \langle z, \hat{\gamma}\rangle$ *ist ein erwartungstreuer linearer Schätzer für* $\Gamma(\gamma)$.

(ii) Die Varianz von T_Γ *ist* $\sigma^2||a||^2$, *wobei* $a = A\big(A^\top A\big)^{-1} z$.

(iii) Ist $S : \mathbb{R}^n \to \mathbb{R}$ ein weiterer linearer und erwartungstreuer Schätzer[6] für $\Gamma\gamma$, so hat S eine nicht kleinere Varianz als T_Γ. Der Schätzer T_Γ ist auch der einzige Schätzer, bei dem die kleinste Varianz realisiert wird.

Beweis: (i) Da $X \mapsto \hat{\gamma}$ linear ist, ist die Linearität von T_Γ klar. Der Erwartungswert von T_Γ ergibt sich aus Linearitätsgründen zu $\langle z, E(\hat{\gamma})\rangle$, und diese Zahl ist wegen Teil (ii) gleich $\langle z, \gamma\rangle$. T_Γ schätzt also $\Gamma(\gamma)$ erwartungstreu.

(ii), (iii) Sei S wie in (iii) gegeben. Zunächst wählen wir ein $b \in \mathbb{R}^n$ mit $S(X) = \langle b, X\rangle$ für alle x. Nach Voraussetzung ist S erwartungstreu, es gilt also

$$E(S) = E\big(\langle b, A\gamma + \sigma\xi)\rangle\big) = \langle b, A\gamma + \sigma E(\xi)\rangle = \langle b, A\gamma\rangle$$

für alle γ. Außerdem ist, mit $a := \left(\left(A^\top A\right)^{-1} A^\top\right)^\top z$, stets

$$\langle z, \hat{\gamma}\rangle = T_\Gamma(X) = \langle z, \left(A^\top A\right)^{-1} A^\top X\rangle = \langle a, X\rangle;$$

dabei können wir die Formel für a noch zu $a = A\big(A^\top A\big)^{-1} z$ vereinfachen, da $A^\top A$ symmetrisch ist[7].
Es folgt

$$E(T_\Gamma(X)) = E(\langle a, A\gamma + \sigma\xi\rangle) = \langle a, A\gamma\rangle$$

für alle γ. Zusammen mit der Gleichung für den Erwartungswert von S heißt das, dass $b - a$ senkrecht auf dem Bild von A – insbesondere senkrecht auf a – steht, und das impliziert

$$\|b\|^2 = \|a + (b-a)\|^2 = \|a\|^2 + \|b-a\|^2.$$

Es folgt also $\|b\| \geq \|a\|$, und Gleichheit gilt nur im Fall $a = b$.

Nun wollen wir die Varianz ausrechnen, zunächst die von $T_\Gamma(X)$. Der Erwartungswert von $T_\Gamma(X)$ ist $\langle a, A\gamma\rangle$, d.h.

$$\begin{aligned} Var(T_\Gamma(X)) &= E((T_\Gamma(X))^2) - \big(E(T_\Gamma(X))\big)^2 \\ &= E\big(\langle a, A\gamma\rangle + \sigma\langle a, \xi\rangle\big)^2 - (\langle a, A\gamma\rangle)^2 \\ &= \sigma^2 E(\langle a, \xi\rangle)^2 \\ &= \sigma^2 E\big((a^\top \xi)^2\big) \\ &= \sigma^2 E\big((a^\top \xi)(\xi^\top a)\big) \\ &= \sigma^2 \|a\|^2; \end{aligned}$$

dabei haben wir ausgenutzt, dass $E(\xi\xi^\top)$ nach Voraussetzung die Einheitsmatrix ist[8].

[6] Gemeint ist dabei: Wird X gemessen, so soll $S(X)$ der Schätzer von $\Gamma(\gamma)$ sein.
[7] Dann nämlich ist auch $(A^\top A)^{-1}$ symmetrisch.
[8] Für der Identität $a^\top \xi = \xi^\top a$ war nur die Symmetrie des Skalarprodukts wichtig.

Analog zeigt man, dass die Varianz von SX gleich $\sigma^2\|b\|^2$ ist.

Zusammen mit der Rechnung des vorigen Abschnitts folgt, dass T_Γ die bessere Varianz hat und dass Gleichheit nur im Fall $S = T_\Gamma$ gilt. □

Als Korollar beweisen wir noch, dass unser oben eingeführter Schätzer $\hat{\gamma}$ für γ dadurch ausgezeichnet ist, dass bei ihm die kleinste Varianz realisiert wird:

Korollar 4.1.6. *Für jeden erwartungstreuen Schätzer S von γ ist die Varianz nicht kleiner als Spur$\left(A^\top A\right)^{-1}$, der Varianz des Schätzers $\hat{\gamma}$. Nur bei diesem Schätzer wird das Minimum angenommen.*

Beweis: Sei $i \in \{1, \ldots, s\}$ und $e_i \in \mathbb{R}^s$ der i-te Einheitsvektor. Die Abbildung $\Gamma = \langle e_i, \cdot\rangle$ bildet γ auf die i-te Komponente von γ ab. Die beste erwartungstreue Schätzung dafür ist aufgrund des vorigen Satzes $\langle e_i, \hat{\gamma}\rangle$, sie hat – mit $a := A\left(A^\top A\right)^{-1} e_i$ die Varianz

$$\begin{aligned}\sigma^2||a||^2 &= \sigma^2\langle a, a\rangle \\ &= \langle e_i, \left(A\left(A^\top A\right)^{-1}\right)^\top A\left(A^\top A\right)^{-1} e_i\rangle \\ &= \langle e_i, \left(A^\top A\right)^{-1} e_i\rangle,\end{aligned}$$

das ist das i-te Element auf der Diagonalen von $\left(A^\top A\right)^{-1}$. Insbesondere ist die Varianz von $\langle e_i, S\cdot\rangle$ (das ist ein erwartungstreuer Schätzer für γ_i) nicht kleiner als diese Zahl.

Bildet man die Summe über alle i, so ergibt sich, dass die Spur von $(A^\top A)^{-1}$ eine untere Schranke der Varianz von S ist, und es folgt auch, dass das Minimum nur bei $S = A\left(A^\top A\right)^{-1}$ angenommen wird. □

Bemerkung: Wir betrachten noch einmal Beispiel 1 vom Beginn dieses Abschnitts, diesmal in einer allgemeineren Fassung. X sei eine Zufallsvariable mit Erwartungswert μ und Streuung σ. Definiert man $\xi := (X - \mu)/\sigma$, so ist das eine normalisierte „Störung“, Wir schreiben $X = \mu + \sigma\xi$ und haben damit X als lineares Modell interpretiert.

Wie schon bemerkt, ist hier $A = (1, \ldots, 1)^\top$. Man mache sich klar, dass der Satz von Gauß-Markov dann besagt, dass $\overline{X}$ eine erwartungstreue Schätzung für μ und dass die Stichprobenvarianz eine erwartungstreue Schätzung für σ^2 ist. Kurz: Man erhält noch einmal Satz 2.2.3.

Zur Illustration soll noch ein *numerisches Beispiel* ausführlich dargestellt werden. Es geht um ein lineares Problem für den Fall $s = 2$ und $n = 6$: Es werden 6 Größen gemessen, die sich aus den Unbekannten γ_1 und γ_2 linear ergeben.
Hier ist die Designmatrix A:

$$\begin{pmatrix} 4.68 & -3.23 \\ -3.72 & -3.60 \\ -2.66 & 3.59 \\ 4.49 & -1.41 \\ -2.30 & 2.76 \\ -1.54 & -4.24 \end{pmatrix}.$$

Nun werden ein $\gamma = (\gamma_1, \gamma_2)^\top$ und ein σ stochastisch erzeugt (aber vorerst nicht angezeigt). Mit diesen Werten generiert der Computer 8 Mal $A\gamma + \sigma\xi$ (ξ ist eine geeignet normierte Gleichverteilung). Hier das Ergebnis (die 8 X-Werte stehen in den Spalten):

$$\begin{pmatrix} 0.63 & 3.20 & 4.507 & 1.38 & 1.71 & 2.14 & 2.61 & 4.55 \\ -18.79 & -18.65 & -20.75 & -18.34 & -19.53 & -21.33 & -18.49 & -20.23 \\ 3.01 & 1.50 & 2.50 & 1.59 & 5.84 & 1.81 & 4.05 & 2.56 \\ 6.38 & 6.31 & 5.96 & 10.42 & 5.25 & 9.07 & 7.27 & 5.52 \\ 1.61 & 2.33 & 4.68 & -0.14 & 2.32 & 4.18 & 1.60 & -0.34 \\ -19.25 & -18.25 & -16.17 & -15.58 & -19.11 & -17.02 & -16.03 & -17.44 \end{pmatrix}.$$

Damit sollen nun γ und σ geschätzt werden, dafür ist $(A^\top A)^{-1}A^\top$ auszurechnen, das ist der optimale Schätzer:

$$\begin{pmatrix} 0.058 & -0.071 & -0.025 & 0.062 & -0.023 & -0.041 \\ -0.035 & -0.076 & 0.049 & -0.005 & 0.037 & -0.077 \end{pmatrix}.$$

Jede einzelne X-Messung führt zu einer Schätzung für γ. Hier die Ergebnisse:

$$\begin{pmatrix} 2.44 & 2.56 & 2.59 & 2.63 & 2.39 & 2.76 & 2.43 & 2.70 \\ 3.06 & 2.84 & 2.93 & 2.56 & 3.24 & 3.05 & 2.77 & 2.81 \end{pmatrix}.$$

(Zum Beispiel gibt die erste X-Messung Anlass zu der Schätzung $\hat{\gamma} = (2.44, 3.06)^\top$). Da wir mehrere Schätzungen zur Verfügung haben, können wir den Mittelwert bilden. So erhalten wir den Vektor $(2.57, 2.91)^\top$, *das* ist unsere favorisiertes γ-Schätzung. Nachträglich kann der wirkliche Wert verraten werden. Der Computer hatte $\gamma = (2.674, 2.975)^\top$ erzeugt, das Ergebnis ist also gar nicht so schlecht.

Nun zu σ. Wir benötigen die Projektionen der X-Werte auf den Bildraum von A. Man erhält sie, wenn man die Matrix $A(A^\top A)^{-1}A^\top$ auf X anwendet. Die Ergebisse stehen als Spalten in der folgenden Matrix:

$$\begin{pmatrix} 1.52 & 2.80 & 2.68 & 4.04 & 0.72 & 3.05 & 2.44 & 3.57 \\ -20.14 & -19.79 & -20.2 & -19.06 & -20.64 & -21.31 & -19.08 & -20.21 \\ 4.53 & 3.41 & 3.63 & 2.21 & 5.30 & 3.64 & 3.49 & 2.91 \\ 6.63 & 7.48 & 7.53 & 8.20 & 6.18 & 8.09 & 7.03 & 8.17 \\ 2.83 & 1.94 & 2.10 & 1.01 & 3.43 & 2.07 & 2.03 & 1.52 \\ -16.76 & -16.00 & -16.45 & -14.94 & -17.46 & -17.22 & -15.52 & -16.09 \end{pmatrix}.$$

Es ist dann jeweils der quadrierte Abstand zu den X-Vektoren auszurechnen, und das Ergebnis ist durch $n - s$, also in unserem Fall durch 4 zu teilen. Wir erhalten die Zahlen

$$\begin{pmatrix} 3.20 & 2.91 & 3.49 & 3.67 & 1.82 & 2.41 & 0.30 & 3.36 \end{pmatrix}.$$

Ihr Mittelwert ist 2.65, *das* ist unsere Schätzung für das wirkliche σ^2. Das war gleich 2.52, der wahre Wert wurde also bei dieser Simulation überschätzt.

Epilog

Das Thema wird in zwei Richtungen fortgesetzt werden. Erstens würde man es ja gerne genauer wissen: Man hat zwar nun bewiesen, dass man γ durch $\hat{\gamma}$ erwartungstreu schätzen kann, doch wie weit liegt denn γ von $\hat{\gamma}$ entfernt? Zum Beispiel wäre das interessant, wenn man nach einer Messung eine Hypothese der Form $\gamma = \gamma_0$ ablehnen oder nicht ablehnen soll oder einen Konfidenzbereich für γ_0 im $\mathbb{R}^s$ zu einem vorgegebenen Niveau konstruieren möchte.

Dazu lässt sich in dieser Allgemeinheit nicht viel sagen, denn die jeweilige Antwort hängt von der Verteilung der Störgrößen ξ_i ab. Theoretisch ist es kein Problem, für jeden Einzelfall die gesuchten Verteilungen (wenigstens approximativ, evtl. auch durch Simulation) zu bestimmen, das ist aber recht schwerfällig. Deswegen wird das hier angeschnittene Problem quasi in der gesamten Literatur nur für den Spezialfall untersucht, dass die ξ_i normalverteilt sind. So werden wir es hier auch machen, die wichtigsten Ergebnisse finden sich in Abschnitt 4.3.

Und zweitens soll auf eine Fragestellung hingewiesen werden, die man jetzt schon formulieren kann, die aber ebenfalls nach Abschnitt 4.3 vertagt wird, weil erst dann quantitative Verfahren eingesetzt werden können.

> Es sei ein echter Unterraum $\tilde{H}$ des $\mathbb{R}^s$ vorgegeben. Lässt sich aus der X-Messung etwas Sinnvolles zu der Vermutung sagen, ob $\gamma \in \tilde{H}$ gilt?

Das hört sich recht trocken an, die Frage umfasst aber eine Reihe wichtiger Spezialfälle:

- Interessiert man sich dafür, ob vielleicht $\gamma_1 = 0$ ist (und folglich weggelassen werden könnte), so sollte man den Unterraum $\tilde{H} = \{\gamma \mid \gamma_1 = 0\}$ betrachten.

- Sind vielleicht alle γ_i gleich? Dann ist sicher der Unterraum

$$\tilde{H} = \{(x, \ldots, x) \mid x \in \mathbb{R}\} \subset \mathbb{R}^s$$

  von Interesse.

Auch das klingt noch nicht wirklich spannend, berührt aber oft äußerst wichtige Fragen (Gibt es Nebenwirkungen bei einem bestimmten Medikament? Haben zwei zu testende Kopfschmerzmittel die gleiche Wirkung?)

Mal angenommen, es gilt $\gamma \in \tilde{H}$, also $A\gamma \in A\tilde{H} := H$. In diesem Fall wird doch die beste Approximation von X an ein Element von U, also das Element $A\hat{\gamma}$, „nahe an“ H liegen. Umgekehrt: Wenn $A\hat{\gamma}$ „weit weg“ von H liegt, dann ist die Hypothese $\gamma \in \tilde{H}$ sicher zu verwerfen.

Auch für diesen Fall gilt sinngemäß das, was vor wenigen Zeilen gesagt wurde: Praktisch (und in geringem Maß auch theoretisch) wird das auch im Fall beliebiger Verteilungen der ξ_i quantitativ zu behandeln sein, die Lehrbücher (und wir) konzentrieren uns aber auf den Fall normalverteilter Störungen.

4.2 Das lineare Modell 2 (beliebiger Rang der Designmatrix)

Bisher hatten wir immer vorausgesetzt, dass A vollen Rang hat. Manchmal ist das nicht erfüllt:

> Zu testen seien zwei Düngemittel, es soll festgestellt werden, welches von beiden den Ertrag stärker steigert. Bezeichnet man als γ_1 den „Normalertrag“ und mit γ_2 bzw. γ_3 die durch die Düngemittel hervorgerufenen Effekte, so misst man doch $\gamma_1+\gamma_2$ und $\gamma_1+\gamma_3$ (jeweils plus zufällige Störung, und jeweils im Idealfall auf vielen Feldern). Es ist $s = 3$, und die zugehörige Designmatrix hat in der ersten Spalte Einsen und in den Spalten zwei und drei eine Eins und eine Null oder umgekehrt. Die erste Spalte ist also die Summe der letzten beiden, und deswegen ist der Rang gleich Zwei.

4.2.1 Vorbereitungen zur Linearen Algebra

Bevor wir die Theorie aus dem vorigen Abschnitt auf die allgemeinere Situation übertragen, gibt es einige Vorüberlegungen zu Problemen aus der Linearen Algebra.

Erster Exkurs zur Linearen Algebra: Lösungen von Gleichungssystemen

Es sei A eine beliebige $n \times s$-Matrix und $\gamma \in \mathbb{R}^s$. Welche Informationen über γ lassen sich aus $A\gamma$ gewinnen? Bisher war es so, dass $\gamma \mapsto A\gamma$ eine injektive Abbildung war, da konnte man γ aus $A\gamma$ zurückgewinnen. Wir formulieren das Problem so:

> Es sei $\Gamma : \mathbb{R}^s \to \mathbb{R}$ eine lineare Abbildung. Lässt sich dann $\Gamma\gamma$ aus $A\gamma$ ermitteln?

Hat A vollen Rang, so geht das für alle Γ, im Extremfall, wenn A die Nullmatrix ist, geht es nur für $\Gamma = 0$. Wie sieht es aber für beliebige A aus?

Als konkretes Beispiel zum Kennenlernen betrachten wir das System

$$\begin{aligned} \gamma_1 + \gamma_2 &= 1 \\ \gamma_1 + \gamma_3 &= 4. \end{aligned}$$

Die drei Unbekannten $\gamma_1, \gamma_2, \gamma_3$ sind durch diese zwei Gleichungen nicht eindeutig bestimmt, aber man kann zum Beispiel sagen, dass $\gamma_3 - \gamma_2 = 3$ sein muss. Anders ausgedrückt: Im Fall der durch $\Gamma\gamma := \gamma_3 - \gamma_2$ definierten linearen Abbildung ist $\Gamma\gamma$ aus $A\gamma$ rekonstruierbar.

Nach dem Studium ähnlicher und komplizierter Beispiele kommt man zu dem Ergebnis, dass es wohl genau für diejenigen Abbildungen $\Gamma = \langle w, \cdot \rangle$ geht, bei denen das $w^\top$ als Linearkombination der Zeilen von A entsteht. Wirklich gilt:

Satz 4.2.1. *Eine lineare Abbildung $\Gamma : \mathbb{R}^s \to \mathbb{R}$ sei mit einem geeignet gewählten $w \in \mathbb{R}^s$ als $\gamma \mapsto \langle w, \gamma \rangle$ geschrieben. Dann sind äquivalent:*

(i) *$\Gamma\gamma$ kann aus $A\gamma$ rekonstruiert werden.*

(ii) *$w^\top$ ist Linearkombination der Zeilen von A, d.h. w liegt im Bild der Abbildung $A^\top : \mathbb{R}^n \to \mathbb{R}^s$.*

Beweis: Es ist klar, dass (ii) aus (i) folgt. Nun sei umgekehrt ein $\Gamma = \langle w, \cdot \rangle$ gegeben, für das man $\Gamma\gamma$ stets aus $A\gamma$ ermitteln kann. Insbesondere heißt das, dass aus $A\gamma = 0$ stets $\Gamma\gamma = 0$ folgen muss. (Wäre das nämlich nicht der Fall, gäbe es also ein γ mit $A\gamma = 0$ und $\Gamma\gamma \neq 0$, so wären auch alle $A(\lambda\gamma) = 0$ für $\lambda \in \mathbb{R}$. Die $\Gamma(\lambda\gamma)$ durchlaufen aber verschiedene Werte.)

Anders ausgedrückt heißt das: Bezeichnet man für $i = 1, \ldots, n$ mit

$$W_i : \mathbb{R}^s \to \mathbb{R}$$

die lineare Abbildung, die einem γ das innere Produkt von γ mit der i-ten Zeile von A zuordnet, so gilt:

> Der Schnitt der Kerne der W_i ist im Kern von Γ enthalten.

Das aber impliziert nach dem *Kernlemma* der Linearen Algebra (Einzelheiten folgen gleich), dass Γ eine Linearkombination der W_i ist, und das ist eine Umformulierung der Behauptung.

Als Ergänzung erläutern und beweisen wir noch das

> *Kernlemma:* Es sei V ein $\mathbb{R}$-Vektorraum, und
>
> $$f, g_1, \ldots, g_n : V \to \mathbb{R}$$
>
> seien lineare Abbildungen. Dann ist f genau dann Linearkombination der g_i, wenn der Schnitt der Kerne der g_i im Kern von f enthalten ist.

Eine Richtung ist trivial. Sei umgekehrt die Kernbedingung erfüllt. Wir bilden V durch

$$x \mapsto (g_1(x), \ldots, g_n(x))$$

in den $\mathbb{R}^n$ ab. Auf dem Bildraum definieren wir eine Abbildung h nach $\mathbb{R}$ durch die Vorschrift

$$(g_1(x), \ldots, g_n(x)) \mapsto f(x).$$

Das Bemerkenswerte: Wegen der Kernbedingung ist das eine wohldefinierte Abbildung. Dann muss man nur noch folgende Tatsachen kombinieren:

- h ist linear (trivial).
- h kann linear auf den $\mathbb{R}^n$ fortgesetzt werden (das lässt sich das leicht zeigen, zum Beispiel unter Verwendung des Basisergänzungssatzes).
- Jede lineare Abbildung vom $\mathbb{R}^n$ nach $\mathbb{R}$ kann als $\langle a, \cdot \rangle$ mit einem geeigneten $a \in \mathbb{R}^n$ geschrieben werden.

Es ist dann klar, dass $f = a_1 g_1 + \cdots + a_n g$ ist. □

Es ist nicht schwer, das für vektorwertige Funktionen zu verallgemeinern. $A : \mathbb{R}^s \to \mathbb{R}^n$ soll wie bisher vorgegeben sein. Außerdem sei C eine $r \times s$-Matrix, die wir wie üblich mit einer Abbildung vom $\mathbb{R}^s$ in den $\mathbb{R}^r$ identifizieren. Unter welchen Umständen lässt sich $C\gamma$ aus $A\gamma$ ermitteln? Sicher genau dann, wenn alle Komponenten bestimmt werden können, und so ergibt sich

Satz 4.2.2. *Es sei C wie vorstehend. Dann sind die folgenden Aussagen äquivalent:*

(i) $C\gamma$ kann bei bekanntem $A\gamma$ bestimmt werden.

(ii) Es gibt eine $r \times n$-Matrix D, so dass $C = DA$.

Zweiter Exkurs zur Linearen Algebra: Pseudoinverse

Mit *Pseudoinversen*[9] kann man nicht-invertierbare Matrizen „so gut wie möglich“ invertieren. Das wird beim Studium von Designmatrizen mit nicht vollem Rang wichtig werden.

Die Problemstellung

Es sei A eine $m \times n$-Matrix, wir identifizieren sie mit einer Abbildung von $\mathbb{R}^n$ nach $\mathbb{R}^m$. Was lässt sich über die Lösbarkeit der Gleichung

$$Ax = b$$

sagen (b bekannt, x gesucht)? Im Idealfall ist sie stets eindeutig lösbar. Dann ist die zu A gehörige Abbildung bijektiv und es existiert die inverse Matrix/Abbildung A^{-1}. (Das kann – wie bekannt – nur dann gehen, wenn $n = m$ gilt.)

Wie kann man aber „das Beste aus der Situation“ machen, wenn A nicht bijektiv ist?

Angenommen, A ist surjektiv, aber nicht injektiv. Dann gibt es zu b unendlich viele Lösungen x. Genauer: Ist x_0 irgendeine Lösung und k ein Element im Kern von A, so ist auch $x_0 + k$ eine Lösung, und alle Lösungen entstehen so. Anders ausgedrückt: Die Menge der Lösungen bildet einen affinen Unterraum.

Algebraisch sind alle diese Lösungen gleichberechtigt. Um aber doch ein eindeutig bestimmtes Element auszuzeichnen, könnte man dasjenige mit minimaler Norm nehmen; dazu muss der $\mathbb{R}^n$ natürlich als euklidischer Raum aufgefasst werden. Dieses Element minimaler Norm ist dadurch ausgezeichnet, dass es senkrecht auf dem Kern von A steht.

Angenommen, A ist injektiv, aber nicht surjektiv. In diesem Fall wird es für gewisse b gar keine Lösung geben. Das Beste, was man tun kann, ist dann die

[9] Namen und Bezeichnungsweise sind in der Literatur nicht einheitlich. Man findet auch den Begriff „verallgemeinerte Inverse“, und für die Pseudoinverse einer Matrix A gibt es als Bezeichnung neben A^+ (wie hier) auch A' und andere Abkürzungen.

Gleichung $Ax = b$ „so gut wie möglich“ zu lösen. Wenn also „$=b$“ nicht erreicht werden kann, so sollte doch wenigstens der Abstand zwischen Ax und b so klein wie möglich sein. Da $A(\mathbb{R}^n)$ ein abgeschlossener Unterraum des $\mathbb{R}^m$ ist, gibt es so ein Element Ax mit kleinstmöglichem Abstand. *Dieses* x ist sicher der beste Kandidat für die Lösung.

Im *allgemeinen Fall*, wenn also A weder injektiv noch surjektiv ist, kann man beide Ideen kombinieren:

> Unter der *Pseudolösung der Gleichung* $Ax = b$ verstehen wir dasjenige $x_0 \in \mathbb{R}^n$, das zwei Bedingungen erfüllt: Erstens hat Ax_0 minimalen Abstand zu b unter allen Ax, und zweitens hat x_0 minimale Norm unter allen x mit $Ax = Ax_0$.
>
> Diejenige Abbildung, die einem b die Pseudolösung zuordnet, heißt die *Pseudoinverse* zu A. Wir werden sie mit A^+ bezeichnen.

Offensichtlich stimmt die Pseudoinverse mit der „richtigen“ Inversen überein, wenn A bijektiv ist. Welche Eigenschaften hat sie aber im allgemeinen Fall, und wie kann man sie berechnen?

Eigenschaften

Projektionen auf Unterräume werden bei den folgenden Beweisen eine wichtige Rolle spielen. Für einen Unterraum M eines endlichdimensionalen euklidischen Raumes H bezeichen wir wie in Abschnitt 4.1 mit $P_M : H \to H$ die orthogonale Projektion auf M.

Als Erstes zerlegen wir den $\mathbb{R}^n$ in den Kern $\ker A$ von A und das orthogonale Komplement $(\ker A)^\perp$; der $\mathbb{R}^n$ ist die orthogonale Summe aus diesen beiden Unterräumen. Es ist leicht zu sehen, dass die Einschränkung von A auf $(\ker A)^\perp$ eine (natürlich lineare) Bijektion auf das Bild $R(A)$ von A ist[10]. Folglich können wir sie invertieren, diese (wieder lineare) Abbildung soll für die nächsten Zeilen B genannt werden. Erinnert man sich noch an die Konstruktion von A^+, so wissen wir jetzt, dass

$$A^+ = B \circ P_{R(A)},$$

und damit ist der erste Teil des nächsten Satzes, in dem wir einige Eigenschaften von A^+ zusammenstellen, bereits gezeigt.:

Satz 4.2.3.

(i) A^+ *ist eine lineare Abbildung.*

(ii) $A \circ A^+ = P_{R(A)}$.

(iii) $(A^+)^+ = A$.

[10] $R(A)$ steht für „range von A“.

(iv) Es sei $B : \mathbb{R}^m \to \mathbb{R}^n$ linear, und die folgenden beiden Eigenschaften seien erfüllt: Erstens ist $A \circ B = P_{R(A)}$, und zweitens gilt $P_{R(A^\top)} \circ B = B$. Dann ist $B = A^+$.
($A^\top$ bezeichnet die Transponierte von A, sie kann mit einer Abbildung von $\mathbb{R}^m$ nach $\mathbb{R}^n$ identifiziert werden.)

(v) $A^+ = A^\top \circ (A \circ A^\top)^+$.

Beweis: (i) ist schon gezeigt, und (ii) ergibt sich so: Ist $b \in \mathbb{R}^m$ beliebig, so ist doch $P_{R(A)}b$ dasjenige Element y in $R(A)$, das b bestmöglich approximiert; nach Definition von A^+ ist aber auch $A(A^+(b)) = y$, und das beweist die Behauptung.

(iii) Wir analysieren zunächst den Operator A^+. Ein y liegt im Kern genau dann, wenn $0 \in R(A)$ das Element bester Aproximation ist, wenn also $y \in \big(R(A)\big)^\perp$ gilt. Folglich ist der Orthogonalraum des Kerns der Bi-Orthogonalraum von $R(A)$, also gleich $R(A)$.

Um für ein $x \in \mathbb{R}^n$ das Element $A^{++}x$ zu finden, muss man also ein $y \in R(A)$ so wählen, dass A^+y möglichst nahe bei x liegt. Beachte dabei, dass $R(A^+) = (\ker A)^\perp$ gilt.

Sei nun $x \in \mathbb{R}^n$, zunächst nehmen wir $x \in \ker A$ an. Das nächste Element in $(\ker A)^\perp$ ist dann 0, und das bedeutet – da A^+ von $R(A)$ nach $(\ker A)^\perp$ bijektiv wie A^{-1} abbildet –, dass $(A^+)^+x = 0$. Und natürlich ist auch $Ax = 0$.

Ist $x \in (\ker A)^\perp$, so darf man x durch sich selbst optimal approximieren, und es ist $(A^+)^+x = Ax$. Das bedeutet, dass A und $(A^+)^+$ auf zwei Unterräumen übereinstimmen, die den $\mathbb{R}^n$ erzeugen. Da beides lineare Abbildungen sind, müssen sie überall gleich sein.

(iv) Für den Beweis müssen wir uns vorbereitend mit $A^\top$ beschäftigen. Das ist doch diejenige Abbildung von $\mathbb{R}^m$ nach $\mathbb{R}^n$, für die für alle $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ die Gleichung

$$\langle Ax, y\rangle = \langle x, A^\top y\rangle$$

gilt. Das hat sofort zur Folge, dass

$$R(A^\top) = (\ker A)^\perp,$$

man muss nur beachten, dass $x = 0$ der einzige Vektor ist, der auf allen Vektoren senkrecht steht.

Die Strategie ist ähnlich wie im vorigen Beweisteil: Wir zeigen die Gleichheit von A^+ und B bei den $y \in R(A)^\perp$ und den $y \in R(A)$.

Sei zunächst $y \in (R(A))^\perp$. Dann ist $P_{R(A)} = 0$, und die erste der vorausgesetzten Gleichungen liefert $By \in \ker A$, Die zweite Gleichung sagt aber, dass By in $R(A^\top) = (\ker A)^\perp$ liegt. Also muss $By = 0$ gelten. Andererseits ist nach Definition von A^+ auch $A^+y = 0$, in diesem Fall sind wir also schon fertig.

Und nun sei $y \in R(A)$. Dann ist (wegen der ersten Gleichung) $ABy = y$. Nach Definition von A^+ ist A^+y derjenige eindeutig bestimmte Vektor x in $(\ker A)^\perp$, für den $Ax = y$ gilt. Um also $A^+y = By$ zu zeigen, reicht es

nachzuprüfen, dass By die Eigenschaften dieses x hat. Die zweite Gleichung liefert (wieder wegen $R(A^\top) = (\ker A)^\perp$), dass By in $(\ker A)^\perp$ liegt. Und dass $A(By) = y$ ist, wurde schon bemerkt.

(v) Auch hierfür brauchen wir eine Vorbereitung: Wir behaupten, dass $R(A) = R(AA^\top)$ gilt. „$\supset$“ gilt offensichtlich, es bleibt „$\subset$“ zu zeigen. Sei also Ax ein beliebiges Element im Bild, wir können es als Ax_2 mit $x_2 \in (\ker A)^\perp$ schreiben: Zerlege nämlich $x = x_1 + x_2$ in Elemente aus dem Kern und dem orthogonalen Komplement. Nun ist aber wegen $R(A^\top) = (\ker A)^\perp$ der Vektor x_2 als $A^\top y$ schreibbar. Das zeigt $Ax = Ax_2 = AA^\top y \in R(A \circ A^\top)$.

Es beginnt der Hauptbeweis. Wir wollen zeigen, dass für $B := A^\top \circ (A \circ A^\top)^+$ die beiden Bedingungen aus (iv) erfüllt sind. $A \circ B = P_{R(A)}$ gilt wegen (ii) (angewendet auf die Abbildung $AA^\top$) und der Vorbereitung:

$$\begin{aligned} A \circ B &= A \circ A^\top \circ (A \circ A^\top)^+ \\ &= P_{R(A \circ A^\top)} \\ &= P_{R(A)}. \end{aligned}$$

Der zweite Teil ist einfacher einzusehen: Ist M ein Unterraum und gilt $x \in M$, so ist $P_M x = x$. Deswegen gilt, da in B als letztes die Abbildung $A^\top$ angewendet wird, $P_{R(A^\top)} \circ B = B$. □

Eine Berechnungsmöglichkeit

Wenn A invertierbar ist, lässt sich $A^+ = A^{-1}$ natürlich sehr leicht mit den üblichen Verfahren bestimmen. Durch den vorigen Satz ist es möglich, alles auf den Fall von Diagonalmatrizen zurückzuführen:

`Schritt 1:` Es sei A eine quadratische Diagnalmatrix, die Einträge auf der Diagonale seien $\lambda_1, \ldots, \lambda_k, 0, \ldots, 0$ mit von Null verschiedenen λ_κ. Es ist dann leicht zu sehen, dass A^+ die Diagonalmatrix mit den Diagonaleinträgen

$$1/\lambda_1, \ldots, 1/\lambda_k, 0, \ldots, 0$$

sein muss.

> Bezeichne mit B diese Diagonalmatrix. Beachtet man, dass $A = A^\top$ gilt und dass $R(A)$ der Raum
>
> $$\{(x_1, \ldots, x_k, 0, \ldots, 0) \mid x_1, \ldots, x_k \in \mathbb{R}\}$$
>
> ist, so ist offensichtlich, dass die beiden Bedingungen aus Teil (iv) des vorigen Satzes erfüllt sind.

`Schritt 2:` Ist A quadratisch und symmetrisch, so gibt es nach dem Satz von der Hauptachsentransformation eine orthogonale Matrix O, für die $C := OAO^{-1}$ eine Diagonalmatrix ist. C^+ kann nach dem ersten Schritt ausgerechnet werden, und es ist $A^+ = O^{-1}C^+O$.

> Das kann wieder leicht mit Teil (iv) des Satzes nachgeprüft werden. Man muss nur vorbereitend zeigen, dass $P_{R(C)} = OP_{R(A)}O^{-1}$ gilt.

`Schritt 3:` Ist nun A ganz beliebig, so können wir A^+ mit Teil (v) des Satzes ausrechnen. Danach ist $A^+ = A^\top \circ (A \circ A^\top)^+$, und kann $(A \circ A^\top)^+$ nach Schritt 2 bestimmt werden, denn $A \circ A^\top$ ist eine quadratische symmetrische Matrix.

4.2.2 Schätzbare Aspekte

Wieder geht es um das lineare Modell

$$X = A\gamma + \sigma\xi,$$

diesmal wird aber nicht vorausgesetzt, dass A vollen Rang hat. Dann besteht natürlich keine Hoffnung, aus einer Stichprobe X Informationen über die Komponenten von γ zu erhalten, denn unsere Messung unterscheidet nicht zwischen γ und $\gamma + \tau$ für beliebige τ im Kern von A.

Deswegen muss man mit weniger zufrieden sein. Man betrachtet $r \times s$-Matrizen C und hofft, Informationen über $C\gamma$ aus X zu gewinnen. $C\gamma$ heißt manchmal ein *Aspekt* von γ.

Definition 4.2.4. *Gegeben sei eine $r \times n$-Matrix S. Wir interpretieren sie als Schätzer für $C\gamma$: Der Vektor SX ist unsere Schätzung für $C\gamma$.*

S heißt ein erwartungstreuer Schätzer für *C, wenn für alle γ der Erwartungswert von SX gleich $C\gamma$ ist*[11)]*:*

$$E(SX) = C\gamma.$$

Man sagt dann auch, dass C schätzbar *ist.*

Gibt es so etwas? Aufgrund unserer Charakterisierung identifizierbarer Aspekte lässt sich schnell ein Kriterium finden:

Satz 4.2.5. *Die folgenden Aussagen sind äquivalent:*

(i) Es gibt einen linearen erwartungstreuen Schätzer S für C.

(ii) Es gibt eine $r \times n$-Matrix D, so dass $C = DA$ gilt.

(iii) $C\gamma$ kann aus $A\gamma$ ermittelt werden.

(iv) $C = CA^+A$.

Beweis: Der Erwartungswert von SX ist doch gleich

$$\begin{aligned} E(SX) &= E\big(S(A\gamma + \sigma\xi)\big) \\ &= SA\gamma + \sigma SE(\xi) \\ &= SA\gamma. \end{aligned}$$

[11)] Manchmal liest man „erwartungstreuer Schätzer für $C\gamma$“, doch das ist missverständlich, da man die Forderung auf ein spezielle γ beziehen könnte.

Das ist genau dann stets gleich $C\gamma$, wenn $C = SA$ gilt, und damit ist gezeigt, dass (ii) aus (i) folgt. Die Umkehrung gilt wegen der gleichen Rechnung trivialerweise, und die Äquivalenz von (ii) und (iii) wurde in Satz 4.2.2 bewiesen.

Klar ist auch, dass (ii) eine Folgerung aus (iv) ist. Sei schließlich (ii) vorausgesetzt. Wir zeigen, dass die Gleichung $C\gamma = CA^+A\gamma$ auf den Unterräumen $\ker A$ und $(\ker A)^\perp$ gilt, aus Linearitätsgründen gilt sie dann auf dem ganzen $\mathbb{R}^s$. Für $\gamma \in \ker A$ sind beide Seiten 0, hier wird (ii) ausgenutzt. Und auf $(\ker A)^\perp$ ist $\gamma = A^+A\gamma$ nach Konstruktion von A^+, d.h. es gilt $C\gamma = CA^+A\gamma$. □

Wir nehmen einmal an, dass zu C ein linearer erwartungstreuer Schätzer existiert. In der Regel wird es dann viele solche Schätzer geben. Als Vorbereitung zum Hauptsatz dieses Abschnitts (Satz 4.2.7) zeigen wir schon hier, dass der Schätzer CA^+ eine gewisse Optimierungsaufgabe löst[12)].

Lemma 4.2.6. *C sei schätzbar, wir bezeichnen mit $\mathcal{M}_C$ die Menge*

$$\{S \mid C = SA\}$$

der C-Schätzer, und wir definieren $\varphi : \mathcal{M}_C \to \mathbb{R}$ durch

$$S \mapsto \mathit{Spur}\, S^\top S.$$

Dann gilt:

(i) φ nimmt bei $S = CA^+$ das Minimum an.

(ii) Für $S \in \mathcal{M}_C$ mit $S \neq CA^+$ ist $\varphi(S) > \varphi(CA^+)$.

Beweis: Die Beweisstrategie ist ähnlich wie im Beweis von Satz 4.1.4(vi). Wir nehmen zunächst an, dass das Bild von A der Raum $\{(x_1, \ldots, x_{\tilde{s}}, 0, \ldots, 0) \mid x_i \in \mathbb{R}\}$ ist[13)]. Jede lineare Abbildung $S : \mathbb{R}^n \to \mathbb{R}^r$ kann dann durch eine Matrix $(D\ F)$ beschrieben werden, wobei D eine $\tilde{s} \times r$-Matrix und F eine $(n - \tilde{s}) \times r$-Matrix ist. Es ist $\mathrm{Spur}\, S^\top S = \mathrm{Spur}\, D^\top D + \mathrm{Spur}\, F^\top F$.

Nun liegt S auf dem Bild von A wegen der Erwartungstreue fest: Es ist $S(A\gamma) = C\gamma$. Folglich ist die Matrix D für alle S die gleiche. Eine minimale Spur für $S^\top$ erhält man damit genau dann, wenn die Spur von $F^\top F$ minimal ist. Dass ist für $F = 0$ der Fall, in allen anderen Fällen ist sie größer[14)]. Nun ist nur noch zu beachten, dass zu $S = CA^+$ die Matrix $F = 0$ gehört.

Im allgemeinen Fall muss man von A zu OA übergehen, wobei O eine orthogonale Matrix ist, die das Bild von A in die ersten Koordinaten dreht. Dann

[12)] Als Problem der Linearen Algebra aufgefasst sieht die Fragestellung recht willkürlich aus. Beim Finden optimaler Schätzer wird es aber eine ganz wichtige Rolle spielen.

[13)] $\tilde{s}$ ist die Dimension von $R(A)$.

[14)] Denn $F^\top F$ ist eine positiv definite Matrix: Nur dann ist die Spur – die Summe der Eigenwerte – Null, wenn alle Eigenwerte Null sind. Und das tritt bei symmetrischen Matrizen nur für die Nullmatrix ein. Weiß man aber, dass $F^\top F = 0$ ist, so muss auch $F = 0$ sein, denn

$$||Fx||^2 = \langle Fx, Fx\rangle = \langle x, F^\top Fx\rangle.$$

kann das vorige Ergebnis für den Spezialfall angewendet werden, und man muss ausnutzen, dass für eine Matrix S die Matrizen $S^\top S$ und $(SO)^\top SO$ die gleiche Spur haben: Das liegt wieder an der Spurgleichung $\operatorname{Spur} AB = \operatorname{Spur} BA$ und der Tatsache, dass $O^\top O$ die Einheitsmatrix ist. □

Bemerkung: Das Lemma hat eine interessante Interpretation in der Sprache der Funktionalanalysis, wir beschränken uns hier auf den endlichdimensionalen Fall.

> Sind H_1 und H_2 endlichdimensionale euklidische Räume und $B : H_1 \to H_2$ eine lineare Abbildung, so versteht man under der *Hilbert-Schmidt-Norm* von B die Wurzel aus der Summe der Eigenwerte von $B^\top B$. Unser Lemma besagt dann: Ist H_1 Unterraum eines euklidischen Raums H_3, so kann B auf genau eine Weise so auf H_3 fortgesetzt werden, dass sich die Hilbert-Schmidt-Norm nicht vergrößert. Diese Fortsetzung ist dadurch gegeben, dass man zunächst H_3 orthogonal auf H_1 projiziert und dann B anwendet.

4.2.3 Designmatrix mit beliebigem Rang: Schätzen der Parameter

Es folgt das sicher weitestgehende Ergebnis dieses Abschnitts:

Satz 4.2.7. *(Satz von Gauß-Markov, allgemeine Designmatrix) Es sei C eine schätzbare $r \times s$-Matrix.*

(i) *$X \mapsto CA^+X$ ist ein linearer erwartungstreuer Schätzer für C.*

(ii) *Die Kovarianzmatrix zu dieser vektorwertigen Abbildung ist $\sigma^2 CA^+(A^+)^\top C^\top$.*

(iii) *Die Varianz des Schätzers, also der Erwartungswert von*

$$||CA^+X - C\gamma||^2,$$

ist $\sigma^2 Spur\left((A^+)^\top C^\top CA^+\right)$.

(iv) *Es ist der beste lineare erwartungstreue Schätzer: Ist B eine $s \times n$-Matrix und $X \mapsto BX$ ein erwartungstreuer Schätzer für C, so ist die Varianz dieses Schätzers gleich $\sigma^2 Spur(B^\top B)$. Diese Zahl ist nicht kleiner als die Varianz von CA^+, und Gleichheit ist nur im Fall $B = CA^+$ zu erwarten.*

(v) *Sei U das Bild von A, die Dimension dieses Unterraums bezeichnen wir mit $\tilde{s}$. Dann gilt:*

$$V^* := \frac{\|X\|^2 - \|P_U X\|^2}{n - \tilde{s}} = \frac{\|X - P_U X\|^2}{n - \tilde{s}}$$

ein erwartungstreuer Schätzer für σ^2 ist.

Bemerkung: Es ist zu betonen, dass auch in diesem Fall (beliebiger Rang von A) die Projektion P_U explizit angegeben werden kann: Man muss nur an Satz 4.2.3(ii) erinnern.

Beweis: (i) Wir müssen den Erwartungswert von

$$\begin{aligned} CA^+X &= CA^+(A\gamma + \sigma\xi) \\ &= CA^+A\gamma + \sigma CA^+\xi \\ &= C\gamma + \sigma CA^+\xi \end{aligned}$$

bestimmen. Da $\xi \mapsto CA^+\xi$ linear ist und $E(\xi) = 0$ gilt, ergibt sich $C\gamma$.

(ii) Die Kovarianzmatrix der vektorwertigen Zufallsvariablen $\sigma D\xi$ ist immer die Matrix $\sigma^2 DD^\top$; das wurde schon im Beweis von Satz 4.1.4(iii) ausgenutzt. Dieses Ergebnis ist hier für $D = CA^+$ anzuwenden.

(iii) Für beliebige D ist die Varianz der Zufallsvariablen $D\xi$ durch $\sigma^2 \text{Spur}\, D^\top D$ gegeben (vgl. den Beweis von Satz 4.1.4(iv)). Hier ist $D = CA^+$ von Interesse, und so ergibt sich die Behauptung.

(iv) Das ist ein Folgerung aus Lemma 4.2.6. Danach ergibt sich für $D = CA^+$ (und auch nur für *diese* Matrix) die kleinsten Spur von $D^\top D$ unter allen Schätzern von C.

(v) Dieser Beweis ist wortwörtlich wie der von Satz 4.1.4(v). □

4.3 Mehrdimensionale Normalverteilungen

Die Ergebnisse der vorigen Abschnitte lassen besonders übersichtliche Interpretationen zu, wenn die auftretenden Störungen normalverteilt sind. Deswegen sollen zunächst die Untersuchungen über die Normalverteilung aus Abschnitt 2.5 fortgesetzt werden.

Definition 4.3.1. *Es seien $Y_1, \ldots, Y_n : \Omega \to \mathbb{R}$ Zufallsvariable. Sie heißen* gemeinsam normalverteilt*, wenn es unabhängige $N(0,1)$-verteilte Zufallsvariable $W_1, \ldots, W_n$, einen Vektor $\mu = (\mu_1, \ldots, \mu_n)^\top$ und eine invertierbare Matrix B so gibt, dass punktweise auf Ω die Gleichung*

$$(Y_1, \ldots, Y_n)^\top = B(W_1, \ldots, W_n)^\top + \mu$$

gilt[15)].

Beispiele:

1. Die Zufallsvariablen $W_1 + 5W_2, W_2$ sind sicher gemeinsam normalverteilt. Man beachte, dass sie nicht unabhängig sind.

[15)] In manchen Büchern wird nicht verlangt, dass B invertierbar ist.

2. Aus bekannten Ergebnissen aus der elementaren Stochastik über die Normalverteilung folgt, dass jedes einzelne Y_k normalverteilt ist. Die Umkehrung gilt nicht: Eine Familie von normalverteilten Zufallsvariablen muss nicht gemeinsam normalverteilt sein.

> Richtig aber ist: Sind die $Y_1, \ldots, Y_n$ linear unabhängig als Funktionen und ist jede Linearkombination normalverteilt, so sind sie gemeinsam normalverteilt.

Alle interessanten Informationen sind in B und μ enthalten:

Satz 4.3.2. *Die $Y_1, \ldots, Y_n$ seien gemeinsam normalverteilt, B und μ seien wie in der Definition. Wir setzen $C := BB^\top = (c_{ij})$.*

(i) Das von $Y = (Y_1, \ldots, Y_n)^\top$ auf dem $\mathbb{R}^n$ induzierte Wahrscheinlichkeitsmaß hat die Dichtefunktion

$$\begin{aligned}\Phi_{\mu,C}(y) &:= \frac{1}{\sqrt{(2\pi)^n |\det C|}} \exp\bigl(-(y-\mu)^\top C^{-1}(y-\mu)/2\bigr) \\ &= \frac{1}{\sqrt{(2\pi)^n |\det C|}} \exp\bigl(-\bigl\|B^{-1}(y-\mu)\bigr\|^2/2\bigr);\end{aligned}$$

das bedeutet, dass die Wahrscheinlichkeit, Y in einer Borelmenge Δ zu finden, stets durch das Integral von $\Phi_{\mu,C}$ über Δ gegeben ist.

(ii) Es ist $E(Y_i) = \mu_i$ und die Kovarianz $Cov(Y_i, Y_j)$ ist gleich c_{ij} für alle i, j.

Beweis: Da B invertierbar sein soll, ist auch C invertierbar, die Funktion $\Phi_{\mu,C}$ ist also wirklich wohldefiniert.

Um die Behauptung zu zeigen, ist natürlich wieder wie in Abschnitt 2.5 der Transformationssatz für Gebietsintegrale anzuwenden (vgl. die Abteilung „technische Vorbereitungen“ in diesem Abschnitt). Zunächst ist also die gemeinsame Dichte der $W_1, \ldots, W_n$ auszurechnen. Das ist leicht, denn da diese Zufallsvariablen standard-normalverteilt und unabhängig sind, ergibt sich sofort die Funktion

$$\frac{1}{\sqrt{(2\pi)^n}} \exp(-x^\top x)/2;$$

man beachte, dass man sie mit der neuen Terminologie als $\Phi_{0,E}$ schreiben kann, dabei bezeichnet ab sofort E die $n \times n$-Einheitsmatrix.

Nach dem Transformationssatz hat Y die Dichtefunktion, die y auf

$$\begin{aligned}&\Phi_{0,E}\bigl(B^{-1}(y-\mu)\bigr)\bigl|\det B^{-1}\bigr| \\ &\quad= \frac{1}{\sqrt{(2\pi)^n}} \exp\bigl(-[B^{-1}(y-\mu)]^\top [B^{-1}(y-\mu)]/2\bigr)\bigl|\det B^{-1}\bigr| \\ &\quad= \frac{1}{\sqrt{(2\pi)^n}}\bigl|\det B^{-1}\bigr| \exp\bigl(-[(y-\mu)]^\top C^{-1}[(y-\mu)]/2\bigr)\end{aligned}$$

abbildet; dabei wurde nur $\left(B^{-1}\right)^\top B^{-1} = \left(BB^\top\right)^{-1}$ ausgenutzt. Wenn man noch beachtet, dass $\det B^{-1} = 1/\det B$ und dass $|\det C| = |\det B|^2$, so ist diese Funktion als $\Phi_{\mu,C}$ identifiziert.

Da die W_k Erwartungswert Null haben, ergibt sich wegen der Linearität des Erwartungswerts, dass $E(Y_i) = \mu_i$ für jedes i gilt. Die Covarianz ist bilinear (d.h. linear in beiden Faktoren), und deswegen ist

$$\mathrm{Cov}\,(Y_i, Y_j) = \sum_{k,l} B_{ik} B_{jl} \mathrm{Cov}\,(W_k, W_l).$$

Nach Voraussetzung ist $\mathrm{Cov}\,(W_k, W_l) = \delta_{kl}$, es bleibt also rechts nur die Summe $\sum_k B_{ik} B_{jk}$ übrig, und die hat den Wert c_{ij}. □

Bemerkungen:

1. Die Verteilung hängt also nicht von B, sondern nur von $BB^\top$ ab. Das ist plausibel, z.B. hat doch (W_2, W_1) sicher die gleiche Verteilung wie (W_1, W_2) und $(-W_2, -W_1)$.

2. C^{-1} ist immer eine symmetrische Matrix. Die Hauptachsentransformation garantiert dann, dass C^{-1} in einer geeigneten Darstellung diagonal ist; da C invertierbar ist und die Form $BB^\top$ hat, müssen alle Diagonalelemente positiv sein.

Nimmt man auch noch (nach Verschiebung des Koordinatenursprungs) $\mu = 0$ an, so ist die Dichtefunktion – bis auf Normalisierung von der Form

$$(y_1, \ldots, y_n)^\top \mapsto \exp\bigl(-(\lambda_1 y_1^2 + \cdots + \lambda_n y_n^2)/2\bigr)$$

mit positiven λ_i.

Wollte man die Dichte durch verschiedene Grautöne visualisieren, so würde das im Zweidimensionalen zu einem Ellipsoid und im Dreidimensionalen zu einer Art „Rugbykugel“ führen, die jeweils achsenparallel sind. Dabei entsprechen die i mit den kleinen λ_i den größeren Halbachsen[16]. Dieses Bild kann man sich auch dann machen, wenn C nicht notwendig diagonal ist, allerdings sind die Halbachsen der n-dimensionalen Ellipsoide dann nicht mehr notwendig achsenparallel.

3. Man mache sich klar, dass die Unabhängigkeit der Komponenten damit zusammenhängt, ob die Achsen des eben beschriebenen Ellipsoids achsenparallel sind.

4. Aufgrund der Konstruktion ist C symmetrisch und positiv definit[17]. Jede derartige Matrix C tritt auf diese Weise auf: Für eine geeignete orthogonale Matrix O ist $C = ODO^\top$, wobei D diagonal mit positiven Einträgen ist; zieht

[16] Die λ_i sind nämlich als die Reziproken der Streuungen zu interpretieren.

[17] D.h. es ist $\langle x, Cx\rangle > 0$ für jedes $x \neq 0$; beachte nur, dass $\langle x, Cx\rangle = \langle B^\top x, B^\top x\rangle = \left\|B^\top x\right\|^2$, und $B^\top$ ist nach Voraussetzung invertierbar.

man aus diesen Diagonalelementen die Wurzel, nennt das Ergebnis D_w und definiert $B = OD_w$, so ist

$$BB^\top = (OD_w)(OD_w)^\top = OD_wD_w^\top O^\top = ODO^\top = C.$$

5. Im Fall $n = 1$ ergibt sich – wie zu erwarten – die Dichte einer Normalverteilung. Der eindimensionale Vektor μ ist der Mittelwert, und hat B die Form (β), so ist β^2 die Varianz.

Definition 4.3.3. *C sei eine positiv definite und symmetrische $n \times n$-Matrix, und $\mu \in \mathbb{R}^n$. Dann heißt das durch $\Phi_{\mu,C}$ auf dem $\mathbb{R}^n$ definierte Wahrscheinlichkeitsmaß die* Normalverteilung $N_n(\mu, C)$.
Sie wird auch Gaußverteilung *genannt.*

Wie im Eindimensionalen kann die Normalverteilung auch auf dem $\mathbb{R}^n$ gut behandelt werden, viele damit zusammenhängende Wahrscheinlichkeiten sind explizit angebbar:

Satz 4.3.4. *Es sei A eine invertierbare $n \times n$-Matrix und $\nu \in \mathbb{R}^n$. Betrachte die Abbildung $Z : x \mapsto Ax + \nu$ von $\mathbb{R}^n$ nach $\mathbb{R}^n$. Dann ist das Bildmaß P_Z von einer Verteilung $N_n(\mu, C)$ unter Z ebenfalls normalverteilt: Es ergibt sich die $N_n(A\mu + \nu, ACA^\top)$-Verteilung.*
Kurz: A mal $N_n(\mu, C)$ plus ν gleich $N_n(A\mu + \nu, ACA^\top)$.

Beweis: Wähle W, Y, B und C wie in Satz 4.3.2: Nach dem Satz ist $Y = BW + \mu$ $N_n(\mu, C)$-verteilt, wo $C := BB^\top$. In der jetzt zu beweisenden Aussage geht es um die Verteilung von $Z = AY + \nu$, also von $ABW + A\mu + \nu$. Nach Satz 4.3.2 ist Z $N_n\big(A\mu+\nu, (AB)(AB)^\top\big)$-verteilt. Damit ist bereits alles gezeigt, denn $(AB)(AB)^\top = ACA^\top$. □

Korollar 4.3.5. *Ist O eine orthogonale Matrix und ist X $N_n(\mu, E)$ verteilt, so ist OX $N_n(O\mu, E)$-verteilt. Insbesondere ist die orthogonale Transformation von unabhängigen standard-normalverteilten Zufallsvariablen wieder unabhängig und standard-normalverteilt.*

Bemerkung: Dieses Ergebnis spielte schon im Beweis von Satz 2.5.10 eine wichtige Rolle.

Als weitere Folgerung wollen wir gleich ein bemerkenswertes Korollar zum Korollar formulieren. Vorher soll an den Zusamenhang zwischen *Unkorreliertheit* und *Unabhängigkeit* erinnert werden.

Hier noch einmal die wichtigsten Definitionen und Tatsachen. Es seien X, Y reellwertige Zufallsvariable, für die die Varianz existiert. Der Einfachheit nehmen wir an, dass der Erwartungswert bei beiden gleich Null ist.

- X, Y heißen *unabhängig*, wenn für beliebige Borelmengen A, B die Ereignisse $\{X \in A\}$ und $\{Y \in B\}$ unabhängig sind.

- X, Y heißen *unkorreliert*, wenn der Erwartungswert von $X \cdot Y = 0$ ist.

  Dazu zwei Kommentare. Erstens sollte man sich an die Motivation in der Einleitung zu Abschnitt 1.5 erinnern: Unkorreliertheit bedeutet, dass sich die „Tendenzen“ von X und Y ausgleichen, positiv oder negativ zu sein. Es kommt „in etwa gleich oft“ vor, dass sie das geiche oder verschiedene Vorzeichen haben.
  Und zweitens sollte bemerkt werden, dass Unkorreliertheit in der Sprache der euklidischen Räume gerade bedeutet, dass X und Y senkrecht aufeinander stehen.

- Es ist eines der ersten fundamentalen Ergebnisse der elementaren Wahrscheinlichkeitsrechnung, dass aus der Unabhängigkeit die Unkorreliertheit folgt[18]. Man braucht diese Tatsache, um nachzuweisen, dass sich für paarweise unabhängige Zufallsvariable die Varianzen addieren; daraus gewinnt man leicht das so genannte Wurzel-n-Gesetz[19].

- Bei überraschend vielen Aussagen der Wahrscheinlichkeitstheorie reicht es, statt der Unabhängigkeit die paarweise Unkorreliertheit zu fordern. Das ist zum Beispiel beim starken Gesetz der großen Zahlen der Fall.

Bei gemeinsam normalverteilten Zufallsvariablen gibt es keinen Unterschied zwischen den beiden Begriffen:

Korollar 4.3.6. *Die Zufallsvariablen $Y_1, \dots, Y_n$ seien gemeinsam normalverteilt und paarweise unkorreliert. Dann sind sie auch unabhängig.*

Beweis: Ohne Einschränkung seien alle Erwartungswerte Null und alle Varianzen 1. Es ist

$$Y := (Y_1, \dots, Y_n)^\top = B(W_1, \dots, W_n)^\top$$

mit unabhängigen standard-normalverteilten W_i. Im Raum der quadratintegrablen Funktionen ist dann

$$\begin{aligned} \langle Y_i, Y_j \rangle &= \langle \sum_s b_{is} W_s, \sum_t b_{jt} W_t \rangle \\ &= \sum_s \sum_t b_{is} b_{jt} \langle W_s, W_t \rangle \\ &= \sum_s b_{is} b_{it}. \end{aligned}$$

Nach Voraussetzung ist $\langle Y_i, Y_j \rangle = \delta_{ij}$ (Kroneckersymbol), und das bedeutet, dass B orthonormale Zeilen hat, also eine orthogonale Matrix ist. Nach dem vorigen Korollar sind dann die Komponenten der vektorwertigen Zufallsvariablen Y unabhängig. □

Bisher hatten wir uns um *invertierbare* Transformationen gekümmert. Wir wissen aber noch nichts – zum Beispiel – über die Verteilung des Vektors

$$(W_1 + W_2 - W_3, W_2 + W_3).$$

[18] Die Umkehrung gilt aber nicht.

[19] Die Streuung von $(X_1 + \cdots + X_n)/n$ ist $\sigma/\sqrt{n}$, wenn die X_i unabhängig und identisch verteilt mit Streuung σ sind

Deswegen soll noch eine etwas allgemeinere Variante des vorstehenden Satzes formuliert werden, die man mit den gleichen Methoden beweisen kann: Durch Auffüllen der Matrizen muss alles auf Satz 4.3.4 zurückgeführt werden.

Satz 4.3.7. *Es sei A eine $k \times n$-Matrix mit Rang k, wobei $k \leq n$, weiter sei $\nu \in \mathbb{R}^k$. Betrachte die Abbildung $Z : x \mapsto Ax + \nu$ von $\mathbb{R}^n$ nach $\mathbb{R}^k$. Dann ist das Bildmaß P_Z von einer Verteilung $N_n(\mu, C)$ unter Z ebenfalls normalverteilt: Es ergibt sich die $N_k(A\mu + \nu, ACA^\top)$-Verteilung.*
Kurz: A mal $N_n(\mu, C)$ plus ν gleich $N_k(A\mu + \nu, ACA^\top)$.

Beweis: (Die Einzelheiten der notwendigen Modifikationen findet man im Beweis von Satz 9.5 im Buch von Georgii.) □

Nach diesen Vorbereitungen soll der *Satz von Gauß-Markov für den Spezialfall normalverteilter Störungen* näher untersucht werden. Es empfiehlt sich, vor dem Beweis des nächsten Satzes noch einmal die Ergebisse aus Abschnitt 2.5 zu wiederholen. Wir betrachten also ein lineares Modell $X = A\gamma + \sigma\xi$ und nehmen an, dass ξ $N_n(0, E)$-verteilt ist.

Satz 4.3.8. *γ und σ seien vorgelegt, wie im vorigen Abschnitt bezeichnen wir mit $\hat{\gamma}$ die Schätzung für γ aus X.*

(i) $\hat{\gamma}$ ist $N_s(\gamma, \sigma^2(A^\top A)^{-1})$-verteilt.

(ii) Mit V^* wie in Satz 4.1.4 gilt: $(n-s)V^*/\sigma^2$ ist χ^2_{n-s}-verteilt.

(iii) $\|A\hat{\gamma} - A\gamma\|^2/\sigma^2$ ist χ^2_s-verteilt; beachte, dass $\|A\hat{\gamma} - A\gamma\| = \|\Pi_U X - E(X)\|$ gilt (dabei ist Π_U die Projektion auf das Bild U von A).

Außerdem ist diese Zufallsvariable unabhängig von V^*. Damit folgt, dass $\|A\hat{\gamma} - A\gamma\|^2/(sV^*)$ $F_{s,n-s}$-verteilt ist.

(iv) Sei H ein echter Unterraum von U, also $r := \dim H < s$. Weiter gelte $A\gamma \in H$. X kann dann sowohl auf U als auch auf H projiziert werden, die Projektionen sollen $\Pi_U X$ und $\Pi_H X$ genannt werden.

Dann gilt: $\|\Pi_U X - \Pi_H X\|^2/\sigma^2$ ist χ^2_{s-r}-verteilt und unabhängig von V^*.

(v) Unter den Voraussetzungen von (iv) ist die Fisher-Statistik

$$F_{H,U} := \frac{n-s}{s-r} \cdot \frac{\|\Pi_U X - \Pi_H X\|^2}{\|X - \Pi_U X\|^2} = \frac{\|A\hat{\gamma} - \Pi_H X\|^2}{(s-r)V^*}$$

$F_{s-r,n-s}$-verteilt (vgl. Satz 2.5.6).

Beweis: (i) Im allgemeinen Satz von Gauß-Markov (Satz 4.1.4) haben wir die explizite Form von $\hat{\gamma}$ angegeben:

$$\hat{\gamma} = \left(A^\top A\right)^{-1} A^\top X.$$

Außerdem kennen wir die Verteilung von $X = A\gamma + \sigma\xi$: Die Störung ξ ist $N_n(0, E)$-verteilt, deswegen ist X $N_n(A\gamma, \sigma^2 E)$-verteilt. Und nun ist Satz 4.3.4 anzuwenden, danach ist $\hat{\gamma}$ $N_n(\mu, C)$-verteilt mit

$$\begin{aligned} \mu &= \big(A^\top A\big)^{-1} A^\top (A\gamma), \\ C &= \big(\big(A^\top A\big)^{-1} A^\top\big)\sigma^2 E\big(\big(A^\top A\big)^{-1} A^\top\big)^\top. \end{aligned}$$

Es ist sicher $\mu = \gamma$, und auch C kann man leicht als $\sigma^2(A^\top A)^{-1}$ ermitteln. Das ist gerade die Behauptung.

(ii) Im letzten Teil des Beweises des Satzes 4.1.4 (Gauß-Markov) hatten wir $(n-s)V^*$ als $\sigma^2 \sum_{k=s+1}^{n} \eta_k$ dargestellt, wobei die $\eta_1, \ldots, \eta_n$ aus den $\xi_1, \ldots, \xi_n$ durch eine orthogonale Transformation entstehen. Wegen Korollar 4.3.5 sind die η_i dann also unabhängig und standard-normalverteilt, $(n-s)V^*/\sigma^2$ ist damit die Summe aus den Quadraten von $n-s$ unabhängigen $N(0,1)$-Zufallsvariablen. Nun muss man sich nur noch an Korollar 2.5.4 erinnern.

(iii) Wir erinnern noch einmal an den schon eben verwendeten Teil des Beweises des Satzes von Gauß-Markov: Wir hatten eine Orthogonalmatrix O dadurch konstruiert, dass wir in die Spalten eine Orthonormalbasis des $\mathbb{R}^n$ geschrieben haben; die ersten s Vektoren sollten dabei eine Basis des Bildes von A sein. Der eben erwähnte Zufallsvektor η war durch $\eta = O^\top \xi$ definiert.

Nun betrachten wir $A\hat{\gamma} - A\gamma$. Wegen $A\hat{\gamma} = \Pi_A X$ und $X = A\gamma + \sigma\xi$ ist also $A\hat{\gamma} - A\gamma = \sigma\Pi_A\xi$, der Erwartungswert des Quadrats der Länge *dieses* Vektors soll berechnet werden. Nun wird unter $O^\top$ das Paar orthogonaler Vektoren $\Pi_A\xi, \xi - \Pi_A\xi$ in die Vektoren $\eta_1 + \cdots + \eta_s,\ \eta_{s+1} + \cdots + \eta_n$ transformiert. Da orthonormale Abbildungen isometrisch abbilden, geht es also um den Erwartungswert der Zufallsvariablen $\sigma^2(\eta_1^2 + \cdots + \eta_s^2)$. Nach Korollar 2.5.4 ist klar, dass das $1/\sigma^2$-fache davon χ_s^2-verteilt ist.

Die Unabhängigkeit von V^* folgt gleich mit, denn diese Zufallsvariable war – wie wir gesehen haben – wie $\eta_{s+1}^2 + \cdots + \eta_n^2$ verteilt. Der Zusatz ist damit wieder eine Folgerung aus Korollar 2.5.4.

(iv), (v) Das wird mit den gleichen Techniken bewiesen und soll deswegen hier nicht ausgeführt werden. (Der Beweis kann im Buch von Georgii auf Seite 316 nachgelesen werden.) □

4.4 Schätzen und testen linearer Hypothesen im Fall normalverteilter Zufallsvariable

Und wozu? Die Bedeutung des der Ergebnisse des vorigen Abschnitts besteht darin, dass *Ungewissheit im Zusammenhang mit linearen Modellen quantifiziert* werden kann. Das kann zum Testen von Hypothesen oder zur Konstruktion von Konfidenzbereichen verwendet werden.

Vorbemerkung 1

Zunächst müssen wir uns noch etwas näher mit der mehrdimensionalen Normalverteilung vertraut machen.

Zum Üben betrachten wir eine Familie $\{W_1, \ldots, W_n\}$ von unabhängigen standardnormalverteilten Zufallsvariablen. Wegen Korollar 2.5.4 ist die Zufallsvariable $Z := W_1^2 + \cdots + W_n^2$ χ_n^2-verteilt. Wenn man also ein $\alpha \in]\,0,1\,[$ vorgibt, kann man aus jeder χ^2-Tafel eine Zahl r bestimmen, so dass $Z \leq r$ mit genau Wahrscheinlichkeit $1-\alpha$ gilt. (Zum Beispiel ist $r = 11.07$ für $n = 5$ und $\alpha = 0.05$.)

Geometrisch formuliert heißt das: Mit Wahrscheinlichkeit $1-\alpha$ liegt der Vektor $(W_1, \ldots, W_n)$ in der euklidischen Kugel mit dem Radius $\sqrt{r}$.

Für uns wird eine *Variante* dieser Beobachtung wichtig sein. Wir betrachten eine symmetrische positiv definite $(n \times n)$-Matrix D und einen Vektor ν. Es soll $Y = (Y_1, \ldots, Y_n)^\top$ $N_n(\nu, D)$-verteilt sein.

Wählt man eine Matrix R, so dass $RDR^\top$ die Einheitsmatrix ist, so wird $R(Y-\nu)$ wegen Satz 4.3.4 $N_n(0, E)$ verteilt sein, und dafür kann man die vorstehenden Überlegungen anwenden.

(Falls möglich, ist eine direkte Rechnung technisch etwas einfacher. Ist Y schon als BW gegeben, kann man R als B^{-1} wählen.)

Beispiele:

1. Betrachte $Y = (a_1 W_1, \ldots, a_n W_n)$, wobei $a_1, \ldots, a_n > 0$. Das führt zu einer Matrix D, die auf der Diagonalen die Einträge a_i^2 hat. Wir wählen R als Diagonalmatrix mit Einträgen $1/a_i$ und gelangen zum folgenden Ergebnis: Mit Wahrscheinlichkeit $1-\alpha$ liegt $(Y_1/a_1, \ldots, Y_n/a_n)$ in der euklidischen Kugel mit Radius $\sqrt{r}$; dabei sind α und r wie vorstehend.

2. Diesmal sei $(Y_1, Y_2) = (W_1 + 2W_2, 4W_2)$. Löst man nach den W_i auf, ermittelt also $(W_1, W_2) = (Y_1 - Y_2/2, Y_2/4)$, so folgt: Mit Wahrscheinlichkeit $1-\alpha$ liegt $(Y_1 - Y_2/2, Y_2/4)^\top$ in der euklidischen Kugel mit dem Radius $\sqrt{r}$.

3. Im Fall $(Y_1, Y_2) = (W_1 + 2W_3, -W_1 + W_2)$ muss anders argumentiert werden, da die Übergangsmatrix nicht invertierbar ist. B und $BB^\top$ sehen wie folgt aus:

$$B = \begin{pmatrix} 1 & 0 & 2 \\ -1 & 1 & 0 \end{pmatrix},\ BB^\top = \begin{pmatrix} 5 & -1 \\ -1 & 2 \end{pmatrix}.$$

Als Matrix R kann man

$$R = \frac{1}{3}\begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}$$

wählen. Das heißt: Der Vektor $(Y_1 + 2Y_2, Y_1 - Y_2)/3$ ist $N_2(0, E)$-verteilt, und deswegen kann man zu α ein r angeben, dass er mit Wahrscheinlichkeit $1-\alpha$ in der Kugel mit dem Radius $\sqrt{r}$ liegt[20)].

[20)] Dieser Vektor ist übrigens – explizit geschrieben – der Vektor

$$(-W_1 + 2W_2 + 2W_3, 2W_1 - W_2 + 2W_3)/3.$$

Vorbemerkung 2

Sei $D : \mathbb{R}^s \to \mathbb{R}^n$ linear. Wie sieht $E := \{x \mid \|Dx\| \leq r\}$ aus? Dazu ist zu beachten, dass

$$\|Dx\|^2 = \langle Dx, Dx\rangle = \langle D^\top Dx, x\rangle$$

gilt und dass $D^\top D$ eine symmetrische positive (und bei uns in der Regel auch invertierbare) Matrix ist. Folglich ist die Menge bei geeigneter Basiswahl von der Form

$$\{x \mid a_1x_1^2 + \cdots + a_nx_n^2 \leq r^2\},$$

und das ist ein *Ellipsoid* im $\mathbb{R}^s$.

Konfidenzbereiche

Wir betrachten ein lineares Modell: $X = A\gamma + \sigma\xi$. Es wird X gemessen, und daraus sollen Aussagen für den Vektor γ und/oder für σ hergeleitet werden. Man muss mehrere Fälle unterscheiden.

σ ist bekannt

Es sei α vorgegeben. Nach dem ersten Teil des Satzes 4.3.8 ist $\hat\gamma - \gamma$ gemäß $N_n(0, \sigma^2(A^\top A)^{-1})$ verteilt. Wir wissen dann aufgrund der Vorbemerkungen, dass es ein Ellipsoid E so gibt, dass $\hat\gamma - \gamma$ mit Wahrscheinlichkeit $1 - \alpha$ in E liegt.

Das heißt aber, dass γ mit Wahrscheinlichkeit $1 - \alpha$ in $\hat\gamma - E$ liegt, oder anders ausgedrückt: $\hat\gamma - E$ ist ein *Konfidenzellipsoid* zum Irrtumsniveau α für den Vektor γ.

σ ist unbekannt, γ soll geschätzt werden

In diesem Fall muss das unbekannte σ mit Hilfe von V^* geschätzt werden, dazu soll Teil (iii) des Satzes herangezogen werden.

Danach ist $Z = \|A(\hat\gamma - \gamma)\|^2/(sV^*)$ gemäß $F_{s,n-s}$-verteilt. Ist also α vorgegeben, so kann man mit Tafelhilfe ein r so finden, dass $Z \leq r$ mit Wahrscheinlichkeit $1 - \alpha$ gilt.

Das bedeutet, dass $\|A(\hat\gamma - \gamma)\|^2 \leq r(sV^*)$ mit dieser Wahrscheinlichkeit ist, und da alles außer γ bekannt ist, gibt es wegen Vorbemerkung 2 ein Ellipsoid E, so dass die Normbedingung äquivalent zu $\hat\gamma - \gamma \in E$ ist. Es folgt: $\hat\gamma - E$ ist ein Konfidenzellipsoid für γ [21)].

σ soll geschätzt werden

Das geht leicht mit Teil (ii) des Satzes. Finde bei vorgegebenem α Zahlen r_1 und r_2, so dass eine χ^2_{n-s}-verteilte Zufallsvariable mit Wahrscheinlichkeit $1 - \alpha$ einen Wert in $[\,r_1, r_2\,]$ hat.

Es folgt die X-Messung, daraus können $\hat\gamma$ und V^* berechnet werden. Da $(n - s)V^*/\sigma^2$ mit Wahrscheinlichkeit $1 - \alpha$ in $[\,r_1, r_2\,]$ liegt, heißt das:

$$[\,(n - s)V^*/r_2, (n - s)V^*/r_1\,]$$

[21)] Hier und im vorigen Fall darf man übrigens $-E$ durch E ersetzen, da die Menge E punktsymmetrisch ist.

ist ein Konfidenzintervall für σ^2.

Eine Linearfunktion von γ soll geschätzt werden

Eine lineare Abbildung Γ sei durch Vorgabe von z gegeben. Gesucht ist ein Konfidenzinervall für $\Gamma\gamma = \langle z, \gamma\rangle$.

Die Zufallsvariable $Z := \Gamma\hat{\gamma}$ ist wegen Satz 4.3.7 und Teil (i) des vorigen Satzes normalverteilt mit Erwartungswert $\Gamma\gamma$ und Varianz $\sigma^2 z^\top (A^\top A)^{-1} z$. Folglich ist

$$Z^* := \frac{Z - \langle z, \gamma\rangle}{\sigma\sqrt{z^\top (A^\top A)^{-1} z}}$$

standard-normalverteilt. Z^* hängt nur von $A\hat{\gamma}$ ab und ist deswegen unabhängig von V^*: Beachte, dass für gemeinsam normalverteilte Zufallsvariable Orthogonalität und Unabhängigkeit übereinstimmen.

Nun kommt Satz 2.5.8 ins Spiel: Sind $X, Y_1, \ldots, Y_k$ unabhängig und $N(0,1)$, so ist $W = \sqrt{k}X/\sqrt{Y_1^2 + \cdots + Y_k^2}$ nach der t-Verteilung mit k Freiheitsgraden verteilt. In unserem Fall spielt Z^* die Rolle des X in diesem Ergebnis, es ist $k = n - s$ zu setzen, und $(n-s)V^*/\sigma^2$ spielt die Rolle von $Y_1^2 + \cdots + Y_k^2$. Es folgt eine t-Verteilung mit $n - s$ Freiheitsgraden für

$$\sqrt{n-s}\frac{Z^*}{\sqrt{(n-s)V^*/\sigma^2}} = \frac{\sigma Z^*}{\sqrt{V^*}}.$$

Damit lassen sich nun Konfidenzintervalle finden, sei dazu α vorgegeben. Wähle ein $r > 0$ (Tabelle!), so dass eine t_{n-s}-verteilte Zufallsvariable mit Wahrscheinlichkeit $1 - \alpha$ in $[-r, r]$ liegt. Setze $r' := r\sqrt{z^\top (A^\top A)^{-1} z}$. Und dann folgt durch Rückwärtsrechnen aus den vorstehenden Überlegungen: Mit Wahrscheinlichkeit $1 - \alpha$ liegt $\Gamma(\gamma)$ in

$$\left[\, \langle z, \hat{\gamma}\rangle - r'\sqrt{V^*}, \langle z, \hat{\gamma}\rangle + r'\sqrt{V^*} \,\right],$$

das ist also ein Konfidenzintervall für $\Gamma(\gamma)$.

Tests

Wir wissen schon aus Kapitel 3, dass die Ideen, die zu Konfidenzbereichen führen, auch zu Regeln zum Annehmen oder Verwerfen von Hypothesen uminterpretiert werden können. Deswegen sind hier keine neuen Ideen notwendig.

Schlussbemerkungen:

1. Die letzte Aussage in Satz 4.3.8 kann dazu verwendet werden, die Relevanz von Einflussgrößen abzuschätzen. Um das zu erläutern, betrachten wir das Beispiel der linearen Regression: Ist eine Approximation durch ein Modell der Form $x \mapsto \gamma_0 + \gamma_1 x$ erforderlich, oder reicht eigentlich auch eine Beschreibung der Form $x \mapsto \gamma_0$? In diesem Fall wäre der Satz mit demjenigen Raum H anzuwenden, der als Bild von $\{(\gamma_0, \gamma_1) \mid \gamma_1 = 0\}$ unter A entsteht. So kann mit einer F-Verteilung getestet werden, ob $\gamma_1 = 0$ eine zulässige Hypothese ist.

Als Verallgemeinerung dieser Idee kann man auch entscheiden, wie hoch der Grad einer Interpolation sinnvollerweise sein sollte: linear? quadratisch? kubisch?

In den Anwendungen wird dieses Verfahren auch in viel komplizierteren Situationen ausgenutzt. Man variiert die Parameter so lange, bis man mit möglichst wenigen γ-Werten möglichst kleine Abweichungen erreicht.

2. Wenn man die Ergebnisse auf die oben angegebenen Beispiele anwendet, ergeben sich zum Teil bekannte Resultate. Im ersten Beispiel etwa kommen die bekannten Tests für Mittelwert und Streuung eindimensional normalverteilter Zufallsvariablen heraus. Im Fall der Regression ergeben sich noch explizite Konfidenzintervalle für γ_0 und γ_1.

3. Analysiert man den mathematischen Hintergrund, so stellt sich heraus, dass alle Ergebnisse auf den folgenden Tatsachen beruhten:

- Wenn man von unabhängigen $N(0,1)$-verteilten Zufallsvariablen $\xi_1, \ldots, \xi_n$ ausgeht, so induziert das durch $\omega \mapsto \big(\xi_1(\omega), \ldots, \xi_n(\omega)\big)$ eine vektorwertige Zufallsvariable $\Phi : \Omega \to \mathbb{R}^n$. Alles läuft dann auf die Frage hinaus, wie für Abbildungen $\Psi : \mathbb{R}^n \to \mathbb{R}^m$ die Zufallsvariable $\Psi \circ \Phi$ verteilt ist.

- Um diese Verteilung zu beschreiben, muss man versuchen, sie in Abhängigkeit von Ψ auf Ergebnisse aus der elementaren Stochastik oder aus Abschnitt 2.5 zurückzuführen.

- *Beispiel 1:* Ist $\Psi = A : \mathbb{R}^n \to \mathbb{R}$ linear, so ist $\Psi \circ \Phi$ nach elementaren Ergebnissen normalverteilt.

- *Beispiel 2:* Für lineare $A : \mathbb{R}^n \to \mathbb{R}^m$ gibt es aufgrund der Ergebnisse zu Beginn dieses Abschnitts ähnlich explizite Bechreibungen.

- *Beispiel 3:* Für $\Psi(x) := \|x\|^2$ ist $\Psi \circ \Phi$ gemäß χ^2_n-verteilt. Allgemeiner: Ist $\Psi(x)$ die quadrierte Länge der Projektion von x auf einen s-dimensionalen Unteraum, so ergibt sich eine χ^2_s-Verteilung.

- *Beispiel 4:* Quotienten aus einer linearen Abbildung und einer Norm führen unter geeigneten Unabhängigkeitsvoraussetzungen auf die t-Verteilung.

- *Beispiel 5:* Die F-Verteilung kommt ins Spiel, wenn es um Quotienten von Normquadraten geht. Wieder muss Unabhängigkeit der betrachteten Vektoren vorausgesetzt werden.

4.5 Ein Intermezzo: Über das Schätzen

Hier folgen – als Ergänzung der Untersuchungen in Kapitel 2 – weitere Informationen zum Thema „Schätzen“. Es wird sich zeigen dass für gemeinsam normalverteilte Zufallsvariable besonders weitgehende Aussagen gemacht werden können.

1. Vorbereitungen

Der Begriff „bedingte Erwartung“ ist maßgeschneidert, um grundsätzliche Probleme rund um das Schätzen angemessen behandeln zu können.

a) Information und σ-Algebren

In der modernen Wahrscheinlichkeitstheorie hat sich herausgestellt, dass sich *Information* durch das Auszeichnen einer σ-Algebra beschreiben lässt. Die Situation ist die folgende:

> Es sei $(\Omega, \mathcal{E}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $\mathcal{E}_0$ eine Teil-σ-Algebra von $\mathcal{E}$. „Die Information, die in $\mathcal{E}_0$ enthalten ist, ist bekannt“ soll dann einfach bedeuten, dass man bei Abfrage dieses Wahrscheinlichkeitsraums (ein ω wird gezogen) für jedes $E_0 \in \mathcal{E}_0$ weiß, ob $\omega \in \mathcal{E}_0$ gilt oder nicht.

Das klingt beim ersten Kennenlernen recht kompliziert, es enthält aber die jedem bekannten elementaren Spezielfälle:

- Ist $B \in \mathcal{E}$, so wird bei der Bestimmung von $P(A|B)$ doch auch angenommen, dass $\omega \in B$ bekannt ist. Das entspricht dem Fall $\mathcal{E}_0 = \{\Omega, \emptyset, B, \Omega \backslash B\}$.

- Manchmal ist es so, dass für eine Zufallsvariable X der Wert von $X(\omega)$ bekannt ist, bevor ω gezeigt wird. Diese Information ist in der kleinsten σ-Algebra enthalten, in der X messbar ist.

  > *Bemerkung 1:* Das ist sinnvoll, denn es kommt wirklich nur auf die von X erzeugte σ-Algebra an. Das sieht man zum Beispiel daran, dass X den gleichen Infomationsgehalt hat wie $3 \cdot X$.
  >
  > *Bemerkung 2:* Das vorstehende Beispiel ist übrigens als Spezialfall enthalten, wenn man X als charakteristische Funktion von B wählt.
  >
  > *Bemerkung 3:* Ganz analog kann man vorgehen, wenn die Werte $X_1(\omega), \ldots, X_k(\omega)$ bekannt sind.

- Die extremen Situationen sind sicher $\mathcal{E}_0$=„triviale σ-Algebra“ (= gar keine Information) bzw. $\mathcal{E}_0 = \mathcal{E}$ (volle Information).

b) Bedingte Erwartung

Mal angenommen, man darf auf die durch $\mathcal{E}_0$ codierte Information zurückgreifen. Man möchte nun wissen, wie diese Information unsere Erwartung in Bezug auf eine feste Zufallsvariable Y verändert (Y soll eine Zufallsvariable sein, für die die Erwartung existiert). Schön wäre doch, wenn man für jedes $E_0 \in \mathcal{E}_0$ schnell ausrechnen könnte, wie die Information „$\omega \in E_0$“ unsere Erwartung für Y verändert. Das führt auf die folgende wichtige

Definition 4.5.1. *Eine Funktion $\varphi : \Omega \to \mathbb{R}$ heißt* bedingte Erwartung von Y unter $\mathcal{E}_0$*, wenn die folgenden beiden Bedingungen erfüllt sind:*

(i) φ ist $\mathcal{E}_0$-messbar;

(ii) für $E_0 \in \mathcal{E}_0$ ist $\int_{E_0} \varphi\, dP = \int_{E_0} Y\, dP$. Man beachte: Wenn man die rechts stehende Zahl durch $P(E_0)$ teilt, erhält man die Erwartung von Y unter der Information, dass ω in $\mathcal{E}_0$ liegt. In diesem Sinn würde φ alle möglichen bedingten Erwartungen verschlüsseln.

Als illustrierendes Beispiel betrachten wir die σ-Algebra $\mathcal{E}_0$, die von einer disjunkten Zerlegung

$$\Omega = B_1 \cup \cdots \cup B_k$$

erzeugt wird. In diesem Fall ist $E(Y|\mathcal{E}_0)$ diejenige Funktion, die auf B_κ den Wert $E(Y|B_\kappa)$, also den Wert $\int_{B_\kappa} Y\, dP/P(B_\kappa)$ hat.

c) Existenz der bedingten Erwartung

Es spielt eine fundamentale Rolle, dass bedingte Erwartungen stets existieren und im Wesentlichen eindeutig bestimmt sind. Geanauer gilt:

Satz 4.5.2. *Ist Y eine Zufalllsvariable auf $(\Omega, \mathcal{E}, \mathbb{P})$ mit existierendem Erwartungswert und $\mathcal{E}_0$ eine Teil-σ-Algebra von $\mathcal{E}$, so gilt:*

(i) Es gibt eine bedingte Erwartung φ von Y unter $\mathcal{E}_0$.

(ii) Je zwei bedingte Erwartungen φ, φ' sind höchstens auf einer Nullmenge verschieden.

Man schreibt $E(Y|\mathcal{E}_0)$ für die bis auf Abänderung auf Nullmengen eindeutig bestimmte bedingte Erwartung. Da diese Funktion stets nur unter dem Integral auftritt, ist die Mehrdeutigeit unerheblich.

Beweis: Für die *Existenz* gibt es einen bemerkenswert eleganten Beweis, wenn man den Satz von Radon-Nikodym kennt: Betrachte auf der σ-Algebra $\mathcal{E}_0$ neben der Einschränkung von P auch das Maß

$$\nu : E \mapsto \int_E Y\, dP.$$

Das zweite Maß ist absolutstetig in Bezug auf das erste und kann deswegen als Integral über eine messbare Funktion dargestellt werden.

Die fast sichere *Eindeutigkeit* folgt sofort aus dem folgenden elementaren Ergebnis der Integrationstheorie: Zwei $\mathcal{E}_0$-messbare und integrierbare Funktionen φ, φ' sind genau dann fast sicher gleich, wenn $\int_E \varphi\, dP = \int_E \varphi'\, dP$ für alle $E \in \mathcal{E}_0$ gilt. □

Beispiel 1: Als Wahrscheinlichkeitsraum betrachten wir $[-1, +1]$ mit der Gleichverteilung. Es sei $Y(x) := x$ und $\mathcal{E}_0$ die von $[-1, 0[$, $[0]$, $]0, 1]$ erzeugte σ-Algebra. Als $E(Y|\mathcal{E}_0)$ kann dann diejenige Funktion gewählt werden, die auf den Atomen von $\mathcal{E}_0$ die Werte $-0.5, 0, 0.5$ hat.

Beispiel 2: Erweitert man $\mathcal{E}_0$ im vorigen Beispiel, indem man die Borelschen Teilmengen von $[0, 1]$ dazunimmt, so ist $E(Y|\mathcal{E}_0)$ die Funktion, die auf $[-1, 0[$ den Wert -0.5 hat und auf den anderen x mit Y übereinstimmt.

d) Eigenschaften der bedingten Erwartung

Für die folgenden Überlegungen sind einige Eigenschaften der bedingten Erwartung von fundamentaler Bedeutung. Die Beweise werden hier nur angedeutet.

Lemma 4.5.3. *Es sei Y eine Zufallsvariable mit existierendem Erwartungswert.*

(i) Die Zuordnung $Y \mapsto E(Y|\mathcal{E}_0)$ ist linear.

(ii) Angenommen, Y ist unabhängig von $\mathcal{E}_0$[22]. Dann ist $E(Y|\mathcal{E}_0)$ die konstante Funktion $E(Y)$.

(iii) Ist g $\mathcal{E}_0$-messbar, so gilt $E(g \cdot Y|\mathcal{E}_0) = g \cdot E(Y|\mathcal{E}_0)$.

Beweis: (i) Als Beispiel betrachten wir die Aussage $E(aY|\mathcal{E}_0) = aE(Y|\mathcal{E}_0)$. Wenn man bemerkt hat, dass $aE(Y|\mathcal{E}_0)$ $\mathcal{E}_0$-messbar ist und dass für alle $E_0 \in \mathcal{E}_0$ die Gleichung

$$\int_{E_0} aY\, dP = \int_{E_0} aE(Y|\mathcal{E}_0)\, dP$$

gilt, ist man schon fertig.

(ii) Sei φ die konstante Funktion $E(Y)$, sie ist sicher $\mathcal{E}_0$-messbar. Es ist noch zu zeigen, dass stets $\int_{E_0} Y\, dP = E(Y)P(E_0)$ gilt. Das beweist man zunächst für den Spezialfall, dass Y eine charakteristische Funktion χ_E ist: Dasfür folgt die Aussage sofort aus der vorausgesetzten Unabhängigkeit. Im allgemeinen Fall ist Y durch Treppenfunktionen zu approximieren, mit Hlfe der Linearität des Integrals und der Vertauschbarkeit von Limes und Integration kann der Beweis leicht vervollständigt werden.

(iii) Die rechts stehende Funktion ist sicher $\mathcal{E}_0$-messbar, es muss noch nachgewiesen werden, dass stets $\int_{E_0} gY\, dP = \int_{E_0} g \cdot E(Y|\mathcal{E}_0)$ gilt. Um das einzusehen, fixiere man E_0. Die Aussage wird zunächst für den Fall gezeigt, dass g eine charakteristische Funktion χ_{F_0} ist. In diesem Fall folgt die gewünschte Gleichheit daraus, dass $E(Y|\mathcal{E}_0)$ bedingte Erwartung ist und $E_0 \cap F_0$ zu $\mathcal{E}_0$ gehört. □

Sei nun Y eine Zufallsvariable und $\mathcal{E}_0$ eine σ-Algebra in $\mathcal{E}$. Wie sollte man Y aufgrund der in $\mathcal{E}_0$ enthaltenen Information schätzen? Die Tatsache, dass nur die $\mathcal{E}_0$-Information verwendet werden darf, drückt sich durch die $\mathcal{E}_0$-Messbarkeit aus. (Als weniger abstraktes Beispiel kann man an das Schätzen von Y unter Verwendung der Abfrage von $X_1, X_2 \ldots, X_n$ denken, wobei die X_i Zufallsvariable sind[23]. Hier ist $\mathcal{E}_0$ die von den X_i erzeugte σ-Algebra. Es ist wichtig zu bemerken, dass in diesem Fall die Funktionen, die bezüglich dieser σ-Algebra messbar sind, gerade die Funktionen des Typs $\omega \mapsto g(X_1(\omega), \ldots, X_n(\omega))$ sind.) Das führt zu der folgenden

[22] Das soll bedeuten, dass alle Mengen in der von Y erzeugten σ-Algebra von allen $E_0 \in \mathcal{E}_0$ unabhängig sind.

[23] Wie etwa kann man die Kreditwürdigkeit aufgrund des Jahreseinkommens, des Alters und der Wohngegend schätzen?

Definition 4.5.4. *Ein* Schätzer *für Y unter Verwendung von $\mathcal{E}_0$ ist eine $\mathcal{E}_0$-messbare Zufallsvariable.*

Ist Z so ein Schätzer, so kann man versuchen zu entscheiden, wie „gut“ er ist. Als geeignetes Maß hat sich der mittlere quadratische Abstand erwiesen:

$$||Y - Z||_2 = \big(\int_\Omega \big(Y(\omega) - Z(\omega)\big)^2 \, dP(\omega)\big)^{1/2} = \langle Y - Z, Y - Z\rangle^{1/2};$$

der Raum $L^2(\Omega, \mathcal{E}, P)$ der quadratintegrierbaren Funktionen wird dazu mit dem Skalarprodukt $\langle A, B\rangle := \int_\Omega A(\omega)B(\omega)\, dP(\omega)$ versehen.

Das Hauptergebnis in diesem Zusammenhang ist die Tatsache, dass die bedingte Erwartung der in diesem Sinne bestmögliche Schätzer ist:

Satz 4.5.5. *Es sei Y eine Zufallsvariable und $\mathcal{E}_0$ eine σ-Algebra in $\mathcal{E}$. Wir setzen voraus, dass Y quadratintegrabel ist, dass also $E(Y^2)$ existiert. Dann ist $E(Y|\mathcal{E}_0)$ unter allen quadratintegrablen Schätzern bestmöglich: Ist Z ein beliebiger quadratintegrabler Schätzer, so gilt*

$$||Y - Z||_2 \geq ||Y - E(Y|\mathcal{E}_0)||_2.$$

Beweis: Wir argumentieren über das Skalarprodukt: Wenn wir zeigen können, dass $Y - E(Y|\mathcal{E}_0)$ senkrecht auf $Z - E(Y|\mathcal{E}_0)$ steht, sind wir fertig, denn dann können wir im „rechtwinkligen Dreieck“ zwischen den Punkten $Z, Y, E(Y|\mathcal{E}_0)$ den Satz von Pythagoras anwenden:

$$||Y - Z||^2 = ||Y - E(Y|\mathcal{E}_0)||^2 + ||Z - E(Y|\mathcal{E}_0||^2)(\geq ||Y - E(Y|\mathcal{E}_0)||^2).$$

Wir müssen also nur zeigen, dass

$$\langle Y - E(Y|\mathcal{E}_0), Z - E(Y|\mathcal{E}_0)\rangle = 0$$

gilt. Beim Nachweis spielt Lemma 4.5.3(iii) eine wichtige Rolle: $\mathcal{E}_0$-messbare Funktionen können aus der bedingten Erwartung „herausgezogen“ werden. Das wird hier so angewendet: Weil

- Ω zu $\mathcal{E}_0$ gehört und
- Z $\mathcal{E}_0$-messbar ist, folgt
-

$$\begin{aligned}
\langle Z, E(Y|\mathcal{E}_0)\rangle &= \int_\Omega Z \cdot E(Y|\mathcal{E}_0)\, dP \\
&= \int_\Omega E(Z \cdot Y|\mathcal{E}_0)\, dP \\
&= \int_\Omega Z \cdot Y \, dP \\
&= \langle Z, Y\rangle.
\end{aligned}$$

Ganz genauso folgt, dass

$$\langle Y, E(Y|\mathcal{E}_0)\rangle = \langle E(Y|\mathcal{E}_0), E(Y|\mathcal{E}_0)\rangle,$$

und nach diesen beiden Vorbereitungen ist die Gleichung

$$\langle Y - E(Y|\mathcal{E}_0), Z - E(Y|\mathcal{E}_0)\rangle = 0$$

eine unmittelbare Folgerung aus der Bilinearität des Skalarprodukts.

Wir wissen jetzt, wie man auf optimale Weise Y schätzen kann. Im Spezialfall, wenn $\mathcal{E}_0$ von $X_1, \ldots, X_n$ erzeugt wird, kann man den optimalen Schätzer als $g(X_1, \ldots, X_n)$ schreiben, wobei g eine geeignete messbare Funktion ist.

In vielen Fällen lässt sich g aber nur sehr mühsam bestimmen, und deswegen versucht man, einfacher zu behandelnde Schätzfunktionen zu finden.

Definition 4.5.6. *Ein* linearer Schätzer *von Y aus $X_1, \ldots, X_n$ ist eine Funktion des Typs $\hat{X} = a_1X_1 + \cdots + a_nX_n$.*

Bemerkungen und Beispiele:

1. Formal gesehen ist also *jede* Linearkomination der X_i ein linearer Schätzer. Man möchte aber natürlich durch das Schätzen möglichst viel Information über Y bekommen, und deswegen ist es naheliegend, wie in Abschnitt 2.2 gewisse Güteeigenschaften zu betrachten:

- Ein linearer Schätzer $\hat{X}$ von Y heißt *erwartungstreu*, wenn

  $$E(Y) = E(\hat{X})$$

  gilt.

- Ein erwartungstreuer linearer Schätzer heißt *linearer erwartungstreuer Schätzer von kleinster Varianz*, wenn $\sigma^2(Y - \hat{X}) \leq \sigma^2(Y - \hat{Z})$ für alle linearen erwartungstreuen Schätzer $\hat{Z}$ gilt.

2. Damit es überhaupt erwartungstreue Schätzer geben kann, muss $E(Y)$ Linearkombination der $E(X_i)$ sein. Insbesondere muss mindestens ein X_i mit $E(X_i) \neq 0$ existieren, wenn $E(Y) \neq 0$ ist.

3. Auf Englisch heißt das übrigens

   Best linear unbiased estimator,

und dafür wird die Abkürzung **BLUE** verwendet.

4. Bei quadratintegrablen $Y, X_1, \ldots, X_n$ kann man den besten linearen Schätzer leicht mit geometrischen Methoden finden: Man muss nur ein $\hat{X}$ in der linearen Hülle von $X_1, \ldots, X_n$ finden, das kleinsten Abstand zu Y hat.

5. In vielen Fällen ist die bedingte Erwartung $E(Y|X_1, \ldots, X_n)$ ein bester linearer erwartungstreuer Schätzer. Das Beispiel 1 auf Seite 113 zum Beispiel

entspricht doch der Situation $E(Y|X)$, wobei X die Signumsfunktion ist (mit den Werten $-1, 0, +1$, je nachdem, ob das Argument < 0, $= 0$ oder > 0 ist). Wirklich ist hier $E(X|Y) = 0.5 \cdot X$.

Wir hatten uns schon davon überzeugt, dass der Schätzer $E(Y|X_1, \ldots, X_n)$ immer bestmöglich ist. Er ist übrigens auch erwartungstreu, da $E(Y|\mathcal{E}_0)$ immer den gleichen Erwartungswert wie Y hat.
Im Allgemeinen ist muss sich $E(Y|X_1, \ldots, X_n)$ aber nicht linear aus den X_i kombinieren lassen.

> Ein Gegenbeispiel: Sei $[\,-2, 2\,]$ mit der Gleichverteilung versehen. Y sei die Funktion $x \mapsto x$ und X sei eine Treppenfunktion, die auf den Intervallen $[\,-2, -1\,]$, $]\,-1, 0\,[$, $\{0\}$, $]\,0, 1\,]$ und $]\,1, 2\,]$ jeweils konstant ist und auf jedem dieser Intervalle einen anderen Wert annimmt.
>
> $E(Y|X)$ ist dann diejenige Funktion, die auf den Teilintervallen konstant gleich dem Mittelwert von Y auf diesem Intervall ist. Da das nicht von den konkreten Werten von X abhängt, kann man es leicht einrichten, dass $E(Y|X)$ kein Vielfaches von X ist. (Um das Beispiel „fair“ zu machen, sollte X wie Y auch den Erwartungswert Null haben. Wenn man das nicht beachtet, lassen sich schon viel einfachere Gegenbeispiele angeben.)

Bemerkenswerter Weise braucht man sich im Fall normalverteilter Zufallsvariablen keine großen Gedanken um solche Probleme zu machen. Da kann das Bestmögliche schon mit linearen Schätzern erreicht werden:

Satz 4.5.7. *Es seien $Y, X_1, \ldots, X_n$ gemeinsam normalverteilt mit Erwartungswert Null. Dann ist $E(Y|X_1, \ldots, X_n)$ eine Linearkombination der $X_1, \ldots, X_n$.*

Oder anders ausgedrückt: Der beste lineare Schätzer ist auch unter allen *Schätzern der bestmögliche.*

Beweis: Wir schreiben $(Y, X_1, \ldots, X_n)^\top$ als $B(W_0, \ldots, W_n)$, wobei B vollen Rang hat und die W_i unabhängig und $N(0, 1)$-verteilt sind. Nachdem man von den X_i zu geeigneten Linearkombinationen übergegangen ist, darf man annehmen, dass B die folgende Form hat:

$$\begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ \beta_1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \beta_n & 0 & 0 & \ldots & 1 \end{pmatrix}.$$

(Im $n \times n$-Bereich „rechts unten“ ist B also die Einheitsmatrix.)

Wir erinnern zunächst an Korollar 4.3.6: Unkorrellierte gemeinsam normalverteilte Zufallsvariable sind unabhängig. Wir benötigen hier eine Variante dieser Aussage zur

Vorbereitung des Beweises: Sind $X_0, X_1, \ldots, X_n$ gemeinsam normalverteilt mit Erwartungswert Null und ist X_0 zu allen $X_1, \ldots, X_n$ unkorrelliert, so ist

$$E(X_0|X_1, \ldots, X_n) = 0.$$

Beweis dazu: Nach Übergang zu einer geeigneten Linearkombination der $X_1, \ldots, X_n$ ist $X_0, X_1 \ldots, X_n$ eine orthogonale Familie (Stichwort: Gram-Schmidt). Es liegen also *unabhängie* Zufallsvariable vor. Also ist X_0 unabhängig von der von $X_1, \ldots, X_n$ erzeugten σ-Algebra und deswegen ist $E(X_0|X_1, \ldots, X_n)$ gleich der konstanten Funktion $E(X_0)$, also gleich Null. (Vgl. das entsprechende Ergebnis zur bedingten Erwartung in Lemma 4.5.3.)

Nun definieren wir

$$X_0 := W_0 - \beta_1 W_1 - \cdots - W_n.$$

Auf $X_0, X_1, \ldots, X_n$ können wir die Vorbereitung anwenden, da alle $\langle X_0, X_i \rangle$ gleich Null sind, es ist also

$$E(X_0|X_1, \ldots, X_n) = 0.$$

Weiter gilt sicher $E(X_i|X_1, \ldots, X_n) = X_i$ für $i \geq 1$ sowie

$$\begin{aligned}
W_0 &= (X_0 + \beta_1 X_1 + \cdots + \beta_n X_n)/(1 + \beta_1^2 + \cdots + \beta_n^2) \\
W_1 &= X_1 - \beta_1 W_0 \\
&= X_1 - \frac{\beta_1}{1 + \beta_1^2 + \cdots + \beta_n^2} \\
&\vdots \quad \vdots \quad \vdots \\
W_n &= X_n - \beta_n W_0 \\
&= X_n - \frac{\beta_n}{1 + \beta_1^2 + \cdots + \beta_n^2}.
\end{aligned}$$

Wir wissen auch, dass $Y = \alpha_0 W_0 + \cdots + \alpha_n W_n$ ist, und daswegen ist wegen der Linearität der bedingten Erwartung (und wegen $E(X_0|X_1, \ldots, X_n) = 0$)

$$\begin{aligned}
E(Y|X_1, \ldots, X_n) &= E(\alpha_0 W_0 + \cdots + \alpha_n W_n | X_1, \ldots, X_n) \\
&= \frac{\alpha_0}{1 + \beta_1^2 + \cdots + \beta_n^2}(\beta_1 X_1 + \cdots + \beta_n X_n) + \\
&\quad + \alpha_1 \big(X_1 - \frac{\beta_1}{1 + \beta_1^2 + \cdots + \beta_n^2}(\beta_1 X_1 + \cdots + \beta_n X_n) \big) \\
&\quad + \cdots \\
&\quad + \alpha_n \big(X_n - \frac{\beta_n}{1 + \beta_1^2 + \cdots + \beta_n^2}(\beta_1 X_1 + \cdots + \beta_n X_n) \big).
\end{aligned}$$

Und die rechte Seite ist offensichtlich eine Linearkombination der X_i. □

4.6 Varianzanalyse

Die allgemeinen Ergebnisse der vorigen Abschnitte sollen nun in einem Spezialfall näher studiert werden: Wir wollen einige Ideen aus der *Varianzanalyse*

kennen lernen. Wegen der großen Relevanz ist das ein sehr umfangreiches Gebiet, es kann hier wirklich nur um die ersten Grundlagen gehen.

Das Grundproblem besteht darin, die Bedeutung von Einflussfaktoren zu ermitteln: Sind Chinesen klüger als Eskimos? Hat die Gehaltsgruppe der Eltern Einfluss auf den Bildungserfolg? ...

Wir beginnen mit der *Diskussion eines Beispiels*, dann wird etwas *zur allgemeinen Theorie* gesagt, und anschließend werden *einige Anwendungen* besprochen. Im nächsten Abschnitt wird dann noch die *Kovarianzanalyse* erläutert.

Das Standardbeispiel

Das Standardbeispiel in vielen Lehrbüchern sieht wie folgt aus. Es sollen s Düngemethoden miteinander verglichen werden: Unterscheiden sie sich? Welche ist die beste?

Der experimentelle Aufbau ist klar. Man verschafft sich Felder $F_1, \dots, F_s$, die jeweils in mehrere gleich große Teilbereiche unterteilt werden: Das Feld F_1 in $F_{1,1}, F_{1,2}, \dots, F_{1,n_1}$, F_2 in $F_{2,1}, F_{2,2}, \dots, F_{2,n_2}$, F_s in $F_{s,1}, F_{s,2}, \dots, F_{s,n_s}$. Dann wird Methode G_i auf F_i ausprobiert, auf $F_{i,j}$ wird $X_{i,j}$ geerntet.

Das Problem besteht dann darin, aus den $X_{i,j}$ Rückschlüsse zu ziehen: Welche Abweichungen sind zufällig erklärbar, welche deuten auf wirkliche Unterschiede der Methoden hin?

Die Theorie zur Varianzanalyse[24)]

Das Problem wird dadurch modelliert, dass man die Erträge auf F_i (also bei Verwendung der i-ten Methode) als Zufallsvariable auffasst: $X_{i,j} = \mu_i + \sigma\xi_{i,j}$. Wie üblich nimmt man dabei an, dass die $\xi_{i,j}$ unabhängig und identisch verteilt mit Erwartungswert 0 und Varianz 1 sind. Wir werden die Sprechweise („Felder“, „Ernte-Erträge“, „Düngemethoden“ ...) beibehalten, auch wenn meist völlig andere Probleme damit behandelt werden sollen.

Das kann leicht als lineares Modell interpretiert werden. Die Unbekannten, über die wir nachher Hypothesen testen und von denen wir Konfidenzintervalle berechnen wollen, sind hier die $\mu_1, \dots, \mu_s$. Man muss also nur

$$X = (X_{1,1}, \dots, X_{1,n_1}, X_{2,1}, \dots, X_{2,n_2}, \dots, X_{s,1}, X_{s,n_s})^\top$$

[24)] Für „Varianzanalyse“ gibt es m Englischen die Abkürzung „ANOVA“ ; das heißt „ANalysis Of VAriance“.

setzen und als Designmatrix die spezielle Matrix

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 1 \end{pmatrix}$$

wählen; dabei enthält die erste Spalte n_1 Einsen, die zweite n_2 usw. Mit dieser Definition können die Beobachtungen wirklich zu

$$X = A(\mu_1, \dots, \mu_s)^\top + \sigma\xi$$

zusammengefasst werden, wobei ξ der Vektor der $\xi_{i,j}$ ist. Das „n“ des linearen Modells ist hier die Zahl $n_1 + \cdots + n_s$, und s hat die gleiche Bedeutung wie früher.

Da A recht einfach ist, können alle Matrix-Konstruktionen aus dem Satz von Gauß-Markov konkret ausgeführt werden:

- Ein erwartungstreuer Schätzer für $(\mu_1, \dots \mu_s)$ ist doch $(A^\top A)^{-1} A^\top X$. Nun ist $D := A^\top A$ die Diagonalmatrix mit Diagonaleinträgen $n_1, \dots, n_s$, als Schätzer ergibt sich also $(M_1, \dots, M_s)$, wobei

  $$M_i = \frac{1}{n_i}(X_{i,1} + \cdots + X_{i,n_i})$$

  das Stichprobenmittel der Messungen auf den Feldern $F_{i,1}, \dots, F_{i,n_i}$ ist. (Das war natürlich zu erwarten.)

- Möchte man nur μ_i schätzen, so sollte man sich an Teil (iv) des Satzes von Gauß-Markov erinnern. Wir wenden dieses Ergebnis auf die Abbildung $(\mu_1, \dots, \mu_s) \mapsto \mu_i$ an, es folgt, dass M_i der optimale Schätzer für μ_i ist.

- Durch den Varianzschätzer V^* aus dem Satz von Gauß-Markov kann man sich Aussagen über σ^2 verschaffen: Wie beeinflussen zufällige Faktoren Ernte-Erträge?

Das ist wenig spektakulär, interessanter wird es, wenn man die ξ-Komponenten als normalverteilt annimmt. (Das ist sicher gerechtfertigt, da sich viele Einflüssse überlagern). Da lässt sich dann Satz 4.3.8 anwenden:

- Der Vektor $(M_1, \ldots, M_s)$ ist normalverteilt mit Mittelwert $(\mu_1, \ldots, \mu_s)$ und Kovarianzmatrix $\sigma^2 D^{-1}$ (die Matrix D wurde vor wenigen Zeilen definiert). Das bedeutet, dass die M_i unabhängig sind, und M_i ist $N(\mu_i, \sigma^2/n_i)$-verteilt
- Mit Hilfe von V^* kann man nun nicht nur σ^2 schätzen. Man kann durch Nachschlagen in einer Tabellle der χ^2_{n-s}-Verteilung auch Konfidenzintervalle bestimmen und Hypothesen testen.

Am interessantesten ist es aber, den letzten Teil von Satz 4.3.8 anzuwenden. Mit den dortigen Bezeichnungen betrachten wir speziell den eindimensionalen Raum H, der Bild unter A von

$$H_1 := \{(x_1, \ldots, x_s) \mid x_1 = \cdots = x_s\}$$

ist. Die Aussage „$A(\mu_1, \ldots, \mu_s)^\top \in H$“ ist dann gleichwertig zu $\mu_1 = \cdots = \mu_s$, und es tritt für die Verbesserung der Approximation beim Übergang von H zu U (= Bild von A) die F-Verteilung auf:

> Wenn wirklich $A(\mu_1, \ldots, \mu_s)^\top \in H$ gilt, so wird
>
> $$\frac{n-s}{s-1} \cdot \frac{\|\Pi_U X - \Pi_H X\|^2}{\|X - \Pi_U X\|^2}$$
>
> $F_{s-1,n-s}$-verteilt sein. Das soll dazu ausgenutzt werden, um zu entscheiden, ob die Nullhypothese $\mu_1 = \cdots = \mu_s$ („Kein Unterschied der Güte der Düngemethoden!“) abzulehnen ist.

Die hier auftretenden Größen müssen noch konkret berechnet werden. Die Projektion von X auf U ist doch der Vektor

$$(M_1, \ldots, M_1, M_1, \ldots, M_1, \ldots, M_s, \ldots, M_s),$$

und folglich ist

$$V^* = \frac{1}{n-s}\|X - P_U X\|^2 = \frac{1}{n-s}\sum_{j=0}^{s-1}\sum_{i=1}^{n_i}\big(X_{j,i} - M_j\big)^2.$$

Wir betrachten das rechtwinklige Dreieck, das von den Vektoren X, $P_U X$ und $P_H X$ aufgespannt wird. $P_H X$ ist der Vektor $(M, \ldots, M)$, wobei $M = \sum_{i,j} X_{ij}/n$ das Gesamtmittel bezeichnet. Damit kennen wir die Eckpunkte, und die quadrierten Längen können leicht ausgerechnet werden. Für die Hypotenuse erhalten wir $\sum_{i,j}(X_{i,j} - M)^2$, wenn man diese Zahl durch $(n-1)$ teilt, kommt die *totale empirische Varianz* V^*_{tot} heraus. Eine Kathete hat die Länge $(n-s)V^*$, die andere die Länge $\sum_i n_i(M_i - M)^2$. Die Interpretation: Das ist die *Varianz zwischen den Gruppenmittelwerten*, genauer definiert man

$$V^*_{zdG} := \frac{1}{s-1}\sum_i n_i(M_i - M)^2.$$

Aus dem Satz von Pythagoras folgt dann die so genannte *Streuungszerlegung*:

$$(n-1)V^*_{tot} = (n-s)V^* + (s-1)V^*_{zdG}.$$

Wir fassen zusammen:

> Berechne aus der konkret vorliegenden Stichprobe die Zahlen V^*_{zdG} und V^*. Unter Annahme der Nullhypothese „Es ist $\mu_1 = \cdots = \mu_s$“ ist dann V^*_{zdG}/V^* $F_{s-1,n-s}$-verteilt.
> Folglich kann man bei Vorgabe eines α mit Tafelhilfe ein $r > 0$ so finden, dass die Nullhypothese im Fall $V^*_{zdG} > rV^*$ abzulehnen ist.

Ein Beispiel

Traditionellerweise rechnet man die relevanten Zahlen ökonomisch in einer Tafel, der so genannten ANOVA-Tafel, aus. Am Ende ist aber nur V^*_{zdG}/V^* interessant.

Mal angenommen, es ist $s = 6$ und insgesamt gab es $n = 51$ Experimente. Dann sind die Quantile der $F_{5,45}$-Verteilung von Interesse. Für $\alpha = 0.05$ etwa liest man aus der Tafel: Eine $F_{5,45}$-verteilte Zufallsvariable ist mit Wahrscheinlichkeit 0.95 kleiner gleich 2.31. Damit kann dann anhand der konkreten Werte Stellung zu der Frage genommen werden, ob die Nullhypothese „Alle Mittelwerte sind gleich“ zu verwerfen ist.

Bemerkung: Man beachte, dass V^* ein Schätzer für σ^2 ist. Diese Zahl bleibt also bei großem n beschränkt. V^*_{zdG} verhält sich anders. Wenn wirklich M von einigen M_i verschieden ist, wird V^*_{zdG} mit wachsendem n und festem s immer größer. Anders ausgedrückt: Erwartungsgemäß reagiert das Verfahren umso sensibler auf Abweichungen von der Nullhypothese, je größer die Zahl n ist.

> Falls $s = 2$ ist, falls also nur zwei Methoden zu vergleichen sind, gibt es auch eine *alternative Möglichkeit*. Wir betrachten Zufallsvariable X, Y, die $N(\mu_1, \sigma^2)$- bzw. $N(\mu_2, \sigma^2)$-verteilt sein sollen; dabei ist σ unbekannt. X wird n-mal, Y m-mal abgefragt.
>
> Wenn man dann an $\mu_1 - \mu_2$ interessiert ist, so kann man das auf die Überlegungen der Seite 110 zurückführen („Lineare Funktionen schätzen“). Es folgt: Die Hypothese $\mu_1 = \mu_2$ kann mit einem t-Test behandelt werden. Sie ist zum Niveau α abzulehnen, wenn
>
> $$|M_1 - M_2| > r\sqrt{\Big(\frac{1}{n} + \frac{1}{m}\Big)V^*}.$$
>
> Dabei ist r so gewählt, dass eine t_{n+m-2}-verteilte Zufallsvariable mit Wahrscheinlichkeit α in $[-r, r]$ liegt.
>
> Für mehr als zwei zu vergleichende Verfahren sind diese Überlegungen aber nicht anwendbar.

Mehrfaktorielle Varianzanalyse

Nun wird es etwas komplizierter, es geht um den Vergleich von mehreren Einflussgrößen[25]: Wie beeinflussen Bildung der Eltern und Wohngegend den Schulerfolg? Wie hängt das Krebsrisiko vom Rauchverhalten und vom Alkoholkonsum ab?

Abstrakt geht es also um s_1 Einflussgrößen vom Typ 1 und s_2 Einflussgrößen vom Typ 2. Alle sollen getestet werden, es gibt also $s := s_1 s_2$ Parameterkombinationen. Wir sprechen wieder von „Feldern“ $F_{i,j}$, darauf soll der Einfluss (i, j) auf n_{ij} Teilfeldern getestet werden[26].

Auf den zu i, j gehörigen Feldern werde $X_{i,j,1}, \ldots, X_{i,j,n_{ij}}$ gemessen, wir fassen die $X_{i,j,k}$ zu einem Vektor X zusammen (er hat $\sum_{i,j} n_{ij}$ Dimensionen).

Nun folgen einige *wichtige Definitionen* und *Plausibilitätsüberlegungen*. Im Folgenden ist:

- μ_{ij} der Erwartungswert der Messungen auf Feld F_{ij}.
- $\overline{\mu}$ der Mittelwert der Erwartungswerte, also
$$\overline{\mu} = \frac{1}{s_1 s_2} \sum_{ij} \mu_{ij}$$
- $\overline{\mu}_{i*}$ soll der Mittelwert über die Methode i vom Typ 1 sein:
$$\overline{\mu}_{i*} = \frac{1}{s_2} \sum \mu_{ij}.$$
- Analog wird $\overline{\mu}_{*j}$ als Mittelwert der Methode j vom Typ 2 definiert.

Wichtig sind dann die folgenden Überlegungen:

- Hätte die i-te Methode vom Typ 1 keinen Einfluss, sollte
$$\alpha_i := \overline{\mu}_{i*} - \overline{\mu}$$
gleich Null sein. α_i misst also diesen Einfluss.
- Entsprechend misst
$$\beta_j := \overline{\mu}_{*j} - \overline{\mu}$$
den Einfluß der j-ten Methode vom Typ 2.
- Folglich ist
$$(\mu_{ij} - \overline{\mu}) - \alpha_i - \beta_j$$
der Effekt, der nicht durch Methode i vom Typ 1 oder Methode j vom Typ 2 erklärt werden kann. Diese Zahl heißt der *Wechselwirkungseffekt* zwischen den Methoden i und j und wird mit γ_{ij} bezeichnet.

[25] Im Folgenden werden wir uns auf die Behandlung des Falls *zweier* Einflussgrößen beschränken.

[26] Man könnte zum Beispiel an verschiedene Düngemethoden und verschiedene Saattermine denken.

Trivialerweise gilt dann $\mu_{ij} = \overline{\mu} + \alpha_i + \beta_j + \gamma_{ij}$, und interessant könnten die folgenden Hypothesen sein:

- Alle α_i sind Null: Das heißt, dass die Einflüsse vom Typ 1 keinen Einfluss auf das Endergebnis haben.
- Analog werden alle β_j verschwinden, wenn Typ 2 keine Rolle spielt.
- Die Hypothese „Alle γ_{ij} verwchwinden“ bedeutet, dass sich die Einflüsse vom Typ 1 und vom Typ 2 gegenseitig nicht beeinflussen.

Die zugehörigen Hypothesen können durch die Definition geeigneter Unterräume modelliert werden, im Fall normalverteilter Fehler lassen sich auch Tests finden, die mit F-Verteilungen konzipiert werden.

Kein Einfluss der Faktoren von Typ 1 !

Das heißt doch, dass wir $\alpha_1 = \cdots = \alpha_{s_1} = 0$ annehmen. Ausgedrückt für die μ_{ij}, die einen s-dimensionalen Parameterraum aufspannen, sind das s_1 Einschränkungen. Davon sind allerdings nur $s_1 - 1$ wichtig, denn nach Definition ist $\sum_i \alpha_i = 0$; die α-Bedingungen sind also nicht unabhängig voneinander. Die Hypothese, dass alle α verschwinden, entspricht daher der Hypothese, dass der Parametervektor (μ_{ij}) in einem speziellen $(s - s_1 + 1)$-dimensionalen Unterraum H_1 des $\mathbb{R}^s$ liegt. Durch den letzten Teil von Satz 4.3.8 kann das durch eine F-Verteilung getestet werden. (Genauer: eine $F_{s_1-1,n-s_1s_2}$-Verteilung.)

Die für die Berechnung der Testgröße relevanten Zahlen sind wieder leicht zu ermitteln. Hier das Ergebnis in Rezeptform:

- Man gebe ein Konfidenzniveau vor und mache die fraglichen Tests: Auf Feld $F_{i,j,k}$ wird $X_{i,j,k}$ geerntet.
- Berechne die Zahlen

$$M_{ij} := \frac{1}{n_{ij}} \sum_k X_{ijk}, \; M_{i*} := \frac{1}{n_{i*}} \sum_{jk} X_{ijk};$$

  dabei ist $n_{i*} := \sum_j n_{ij}$ die Anzahl der „Felder“ für die die i-te Methode getestet wurde.
- V^* hat hier die Form

$$V^* = \frac{1}{n - s_1 s_2} \sum_{i,j,k} (X_{ijk} - M_{ij})^2.$$

- Wichtig ist dann noch die (geeignet skalierte) Länge der anderen Kathete. Das führt auf die *empirische Varianz zwischen den Gruppen von Faktor 1*:

$$V^*_{zdG1} := \frac{1}{s_1 - 1} \sum_{ij} n_{ij} (M_{i*} - M)^2.$$

- Wähle nun zum vorgegebenen Fehlerniveau α eine Zahl r, so dass eine $F_{s_1-1,n-s_1s_2}$-verteilte Zufallsvariable mit Wahrscheinlichkeit $1-\alpha$ in $[\,0, r\,]$ liegt. Lehne die Hypothese ab, falls

$$V^*_{zdG1} > r\, V^*.$$

Kein gegenseitiger Einfluss der Faktoren!

Diesmal ist von den s Bedingungen $\gamma_{ij} = 0$ auszugehen, der zugehörige Raum hat die Dimension $s_1 + s_2 - 1$. Völlig analoge Überlegungen führen dann zu folgendem „Rezept“:

- Berechne wie im vorigen Beispiel M, die M_{i*} und die analog definierten M_{*j}.

- Bestimme zusätzlich $V^*_{zdG12} := \sum_{ij} n_{ij}(M_{ij} - M_{i*} - M_{*j} + M)^2$, die *Wechselwirkungsvarianz zwischen den beiden Gruppen*.
- Wähle zu vorgegebenem α ein r, so dass eine $F_{(s_1-1)(s_2-1),n-s_1s_2}$-verteilte Zufallsvariable mit Wahrscheinlichkeit $1-\alpha$ in $[\,0, r\,]$ liegt.
- Lehne die Hypothese „Kein gegenseitiger Einfluss“ ab, falls $V^*_{zdG12} > r\,V^*$.

Als *Beispiel* kann man sich eine Messreihe verschaffen, in der die Reaktionszeit unter mehr oder weniger hohem Alkoholeinfluss und der Einnahme eines Medikaments gemessen wird. Als Ergebnis könnte etwa herauskommen: Das Medikament selbst hat keinen Einfluss, Alkohol hat sehr wohl einen Einfluss, und es gibt einen zusätzlichen Wechselwirkungseinfluss.

4.7 Kovarianzanalyse

Eine typische Problemstellung bei der Kovarianzanalyse sieht so aus:

> Man möchte eine Standard-Varianzanalyse durchführen, zum Beispiel soll der Lernerfolg von Schülern bei der Verwendung unterschiedlicher Englisch-Lehrbücher verglichen werden. Nun stellt man allerdings bei Diskussionen in der Klasse fest, dass die Kinder in ihrer Freizeit sehr unterschiedliche Möglichkeiten haben, die Englischkenntnisse anzuwenden. Manche beschäftigen sich intensiv mit Popmusik, in manchen Elternhäuser werden regelmäßig Englischsprachige Sendungen im Fernsehen angeschaut usw.
>
> Diese unterschiedlichen Ausgangssituationen würden das Ergebnis des Schulbuchvergleichs sicher verfälschen, und deswegen versucht man, das durch einen Vorlauf vor der eigentlichen Varianzanalyse auszugleichen, den störenden Einfluss also irgendwie „herauszurechnen“.

Formal stellt sich das Problem so da. Eigentlich möchte man ein ganz gewöhnliches lineares Modell

$$X = A\gamma + \sigma\xi$$

behandeln. In unserem Beispiel wäre es eine Varianzanalyse, bei der A nur aus Nullen und Einsen besteht. Tatsächlich ist das aber ein zu einfaches Modell, in Wirklichkeit wäre

$$X = A\gamma + B\tilde{\gamma} + \sigma\xi$$

angemessener; dabei ist B eine $n \times \tilde{s}$-Matrix, dadurch wird der Einfluss der $\tilde{s}$ weiteren Einflussgrößen modelliert.

Je nach Problemstellung kann nun eine Schätzung von γ oder $\tilde{\gamma}$ interessanter sein, meist will man nur Informationen über γ erhalten.

Aufgrund der allgemeinen Ergebnisse in Abschnitt 4.2 kann man doch so vorgehen:

- Fasse $X = A\gamma + B\tilde{\gamma} + \sigma\xi$ als gewöhnliches lineares Modell auf. Doch Achtung: Die Designmatrix ist jetzt $A_0 := (A\ B)$ (dazu werden einfach die Matrizen A und B nebeneinander geschrieben), und der Parametervektor ist $\gamma_0 := (\gamma_1, \ldots, \gamma_s, \tilde{\gamma}_1 \ldots, \tilde{\gamma}_{\tilde{s}})$.

- Von diesem Parametervektor interessiert nur der γ-Anteil, also die ersten s Komponenten. Diesen Anteil erhält man dadurch, dass man die $s \times (s+\tilde{s})$-Matrix
$$C := \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \end{pmatrix}.$$
auf γ_0 anwendet.

- C ist sicher schätzbar, wenn wir genügend starke Voraussetzungen für A und B fordern. Und deswegen kann ein optimaler Schätzer mit Satz 4.2.7 gefunden werden: Es ist der Schätzer CA_0^+.

Damit besteht die Hauptaufgabe darin, CA_0^+ auszurechnen.

Wir beginnen mit den technischen Voraussetzungen, wobei wir – um das Ganze übersichtlich zu halten – größte Allgemeinheit nicht anstreben. Wir wollen ab hier voraussetzen, dass A und B jeweils vollen Rang haben (also s bzw. $\tilde{s}$). Auch fordern wir, dass die Bildräume von A und B nur die Null in ihrem Schnitt haben[27)].

Zur Abkürzung schreiben wird U für den Bildraum von A und V für den Bildraum von B und setzen $W := U + V$. Es ist dann W das Bild von A_0, und dieser Unterraum des $\mathbb{R}^n$ ist aufgrund unserer Voraussetzungen die direkte Summem aus U und V.

Nun analysieren wir den Schätzer CA_0^+, als Erstes kümmern wir uns um die Wirkung von A_0^+. Ist X gegeben, so wird X zunächst orthogonal auf W projiziert, dieses Element soll hier w genannt werden. Beachte, dass w auf eindeutige Weise als $u + v$ mit $u \in U$ und $v \in V$ schreibbar ist. Dann wird ein inverses Element γ_0 zu w gesucht. γ_0 ist aus einem γ und einem $\tilde{\gamma}$ zusammengesetzt, und da $A_0\gamma_0$ einerseits gleich w $(= u + v)$ und andererseits gleich $A\gamma + B\tilde{\gamma}$ ist, muss wegen der Eindeutigkeit der Darstellung von w notwendig $u = A\gamma$ und $v = B\tilde{\gamma}$ gelten.

Aber $C\gamma_0$ ist gleich γ, und das bedeutet:

> Um das gesuchte γ zu finden, muss man nur die Gleichung $A\gamma = u$ lösen.

Damit haben wir das Problem in zwei Teilprobleme aufgespalten:

[27)] Das heißt einfach, dass man nicht Teile der Designmatrizen ohne Informationsverlust weglassen darf.

- Finde zu vorgegebenem $w \in U + V$ eine konkrete Zerlegung $w = u + v$.
- Finde γ mit $A\gamma = u$.

Dieses γ ist dann unsere beste Schätzung auf der Grundlage von X.

Lösung von Teil 1 des Problems

Es ist hier nützlich, an einen schon bewiesenen und mehrfach verwendeten Sachverhalt zu erinnern:

- Ist $A : \mathbb{R}^s \to \mathbb{R}^n$, so lässt sich die Projektion auf den Bildraum von A explizit als $P_A = A(A^\top A)^{-1}A^\top$ schreiben.

Damit ist die Projektion auf den zum Bild von A orthogonalen Raum durch $P' := Id - P_A$ gegeben.

Lemma 4.7.1. *Für $x \in \mathbb{R}^n$ sei w die orthogonale Projektion auf W, und w sei als $u + v$ mit $u \in U$ und $v \in V$ geschrieben. Dann ist $v = B(B^\top P'B)^{-1}B^\top P'x$.*

Beweis: Sei $Q := B(B^\top P'B)^{-1}B^\top P'$. Wir behaupten, dass Q erstens wohldefiniert und zweitens die „schiefe“ Projektion von $\mathbb{R}^n$ auf V ist, die U und $W^\perp$ auf 0 abbildet. Dazu ist zu zeigen:

- Wir zeigen, dass $B^\top P'B$ injektiv (und folglich bijektiv) ist. Sei dazu x mit $B^\top P'Bx = 0$ gegeben. Dann liegt $y := P'Bx$ einerseits in $V^\perp$ (da der Kern von $B^\top$ der Orthogonalraum des Bildes von B ist) und andererseits im Bild von P', also in $U^\perp$. Also ist $y \in U^\perp \cap V^\perp = W^\perp$.

  Es ist aber auch $y = Bx - P_U(Bx) \in W$, also $y \in W \cap W^\perp = \{0\}$. Das bedeutet $Bx \in U$, und da $U \cap V = \{0\}$ gilt, ist $Bx = 0$. Da B injektiv ist, heißt das $x = 0$.

- Q ist idempotent, also $Q^2 = Q$: Das ist aufgrund der Definition klar.

- $Qx = x$ für $x \in V$. Wirklich gilt, wenn x von der Form $B\gamma$ ist, dass

$$Qx = B(B^\top P'B)^{-1}B^\top P'B\gamma = B\gamma = x.$$

- Für $x \in U$ ist $Qx = 0$. Das ist klar, denn für diese x ist $P'x = 0$.

- Ist $x \in W^\perp = U^\perp \cap V^\perp$, so ist zunächst (wegen $x \in U^\perp$) $P'x = x$. Andererseits gilt für $x \in V^\perp$, dass $B^\top x = 0$ gilt. (Denn ist

$$\langle x, B\tilde{\gamma}\rangle = \langle B^\top x, \tilde{\gamma}\rangle = 0$$

  für alle $\tilde{\gamma}$, so folgt $B^\top x = 0$). Insgesamt folgt, dass $Qx = 0$ für solche x gilt.

Damit ist das Lemma bewiesen. □

Lösung von Teil 2 des Problems

Dafür brauchen wir nur an schon Bekanntes zu erinnern: γ lässt sich aus einem $u = A\gamma$ leicht durch $\gamma = (A^\top A)^{-1}A^\top u$ rekonstruieren.

Wir fassen zusammen:

Satz 4.7.2. *Um γ optimal aus X zu schätzen, verfahre man wie folgt:*

(i) Berechne zunächst $v := B(B^\top P'B)^{-1}B^\top P'X$.

(ii) Korrigiere X zu $X' := X - v$.

(iii) Schätze dann γ durch $\hat{\gamma} := (A^\top A)^{-1}A^\top X'$.

Bemerkungen und Beispiele:

1. Die Situation ist besonders einfach, wenn U und V senkrecht aufeinander stehen, wenn also

$$\langle A\gamma, B\tilde{\gamma}\rangle = \langle B^\top A\gamma, \tilde{\gamma}\rangle = 0$$

für alle $\gamma, \tilde{\gamma}$ gilt. Es folgt, dass $B^\top A$ die Nullabbildung ist. Damit ist auch $B^\top P_A = 0$, d.h., es ist $B^\top P' = B^\top$. Deswegen vereinfacht sich die Definition von Q zu

$$Q = B(B^\top B)^{-1}B^\top.$$

Das ist aber gerade der optimale Schätzer für $\tilde{\gamma}$ aus X, und wir können zusammenfassen:

> Wenn die Bilder von A und B senkrecht aufeinander stehen, kann die Kovarianzanalyse wie folgt interpretiert werden:
>
> - Gegeben sei die Stichprobe $X = A\gamma + B\tilde{\gamma} + \sigma\xi$.
> - Fasse sie zunächst als lineares Modell $X = B\tilde{\gamma} + (A\gamma + \sigma\xi)$ auf: Die Regressoren sind nur die Komponenten von $\tilde{\gamma}$, und $A\gamma + \sigma\xi$ ist die „Störung“.
> - Schätze nun aufgrund dieses Modells $\tilde{\gamma}$ optimal: Die Schätzung bezeichnen wir mit v.
> - Korrigiere X zu $X - v$ („die Stichprobe wird um den Einfluss der $\tilde{\gamma}$-Parameter bereinigt“) und schätze damit das γ.

2. Wir nehmen das Beispiel vom Beginn dieses Abschnitts noch einmal auf. Zu testen sind zwei Englischbücher, in jeder Gruppe sind 4 Personen beteiligt. Protokolliert sind der Lernerfolg und die geschätzen Vorkenntnisse[28)]:

[28)] In Bezug auf Skalen, die die jeweilige Größe hoffentlich linear wiedergeben ...

Person	Buch	Lernerfolg	Vorkenntnisse
1	1	12	0
2	1	10	1
3	1	14	2
4	1	15	6
5	2	4	2
6	2	7	4
7	2	2	1
8	2	3	0

Das Modell der Kovarianzanalyse besteht hier aus den Matrizen

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \text{ und } B = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 6 \\ 2 \\ 4 \\ 1 \\ 0 \end{pmatrix},$$

und der X-Vektor ist die Spalte „Lernerfolg“.

Im ersten Schritt ignorieren wir die Vorkenntnisse und werten das Modell nur unter Verwendung des A-Modells aus: Das ist eine ganz normale Varianzanlyse.

Erwartungsgemäß ergeben sich als Parameter die Mittelwerte über die Lernerfolge bei Verwendung von Buch 1 bzw. Buch 2, also die Zahlen 12.75 und 4.

Nun sollen die Vorkenntnise berücksichtigt werden. Mit den vorstehenden Bezeichnungen ist in diesem Fall die Projektion P' auf das orthogonale Komplement von Bild A durch

$$\begin{pmatrix} 0.75 & -0.25 & -0.25 & -0.25 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.25 & 0.75 & -0.25 & -0.25 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.25 & -0.25 & 0.75 & -0.25 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.25 & -0.25 & -0.25 & 0.75 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.75 & -0.25 & -0.25 & -0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.25 & 0.75 & -0.25 & -0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.25 & -0.25 & 0.75 & -0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.25 & -0.25 & -0.25 & 0.75 \end{pmatrix}.$$

gegeben, und die „schiefe“ Projektion auf das Bild von B, also $(B^\top P' B)^{-1} B^\top P'$, ist durch

$$\begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.08 & -0.04 & -0.01 & 0.13 & 0.01 & 0.08 & -0.03 & -0.06 \\ -0.15 & -0.08 & -0.02 & 0.25 & 0.02 & 0.15 & -0.05 & -0.12 \\ -0.46 & -0.25 & -0.05 & 0.76 & 0.05 & 0.46 & -0.15 & -0.36 \\ -0.15 & -0.08 & -0.02 & 0.25 & 0.02 & 0.15 & -0.05 & -0.12 \\ -0.31 & -0.17 & -0.03 & 0.51 & 0.03 & 0.31 & -0.10 & -0.24 \\ -0.08 & -0.04 & -0.01 & 0.13 & 0.01 & 0.08 & -0.03 & -0.06 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix}.$$

definiert. Das ergibt einen „Korrekturvektor“

$$v = \begin{pmatrix} 0.00 & 0.79 & 1.58 & 4.73 & 1.58 & 3.15 & 0.79 & 0.00 \end{pmatrix}^\top,$$

und der „bereinigte“ Vektor $X - v$ ist

$$\begin{pmatrix} 12.00 & 9.21 & 12.42 & 10.27 & 2.42 & 3.85 & 1.21 & 3.00 \end{pmatrix}^\top.$$

(Man sieht, dass der Lernerfolg niedriger gewertet wird, wenn es Vorkenntnisse gibt.) *Damit* wird nun das lineare Modell mit der A-Designmatrix ausgewertet, man erhält die Parameter (in unserem Fall die Erfolgszahlen für die beiden Lehrbücher) 10.98 und 2.62.

Alternativ hätte man auch mit der Designmatrix $(A\ B)$ und dem originalen X arbeiten können. Erwartungsgemäß ergeben sich die gleichen Werte für γ_1 und γ_2, und zusätzlich erhält man noch $\gamma_3 = 0.79$. Diese Zahl kann als Einfluss der Vorkenntnisse interpretiert werden.

Kapitel 5

Nichtparametrische Verfahren

In den vorigen Kapiteln wurden Verfahren besprochen, um Zahlen oder Vektoren zu schätzen oder damit zusammenhängende Vermutungen zu testen. Man spricht dann von *parametrischen Verfahren*. Viele wichtige Fragen sind aber dadurch nicht abgedeckt: Sind gewisse Zufallsvariable unabhängig? Verhält sich ein vorgelegter Würfel wirklich so wie behauptet? ...

In diesem Kapitel wollen wir auf einige dieser Probleme eingehen. Meistens wird es um Methoden gehen, die auch bei Vorliegen beliebiger Verteilungen anwendbar sein müssen, die entsprechenden Verfahren heißen *verteilungsfrei*.

In *Abschnitt 5.1* wird die Hypothese getestet, ob eine vorgelegte Wahrscheinlichkeit mit einer speziellen, genau beschriebenen übereinstimmt. Überraschenderweise führt das wieder auf die χ^2-Verteilung, der zugehörige Test heißt der *χ^2-Anpassungstest*. Danach, in *Abschnitt 5.2*, wird es um *χ^2-Tests auf Unabhängigkeit* gehen. Damit kann – einheitlich für alle Verteilungen – getestet werden, ob zwei Zufallsvariable unabhängig sind.

Abschnitt 5.3 ist dann den *Rangtests* gewidmet. Die Hypothese, ob zwei Wahrscheinlichkeitsmaße auf $\mathbb{R}$ gleich sind, kann dabei dadurch getestet werden, dass die Ränge von Stichproben miteinander verglichen werden.

Es folgt dann noch die Besprechung des *Kolmogoroff-Smirnoff-Tests*: Liefern Stichproben einen Anhaltspunkt dafür, ob eine bestimmte kontinuierliche Verteilung vorliegt? Das wird in *Abschnitt 5.4* diskutiert werden.

5.1 Der χ^2-Anpassungstest

Angenommen, jemand behauptet, dass ein Würfel fair ist oder auf ganz bestimmte Weise gefälscht. Wie könnte man das testen? Man wird ihn „genügend oft“ werfen und die Ergebnisse zählen. Sind die empirischen Häufigkeiten „na-

he genug“ bei den behaupteten, ist alles in Ordnung, andernfalls sind Zweifel angebracht. Doch wie oft ist „genügend oft“, wie nahe ist „nahe genug?

Sei ein Wahrscheinlichkeitsraum auf $\{1,\dots,s\}$ durch Vorgabe der Zahlen $p_1,\dots,p_s>0$ definiert. Wir fragen ihn n-mal ab, wobei n groß gegen s sein soll: $h_n(1)$-mal erscheint die 1, $h_n(2)$-mal die 2 usw. Es ist also $h_n(1)+\cdots+h_n(s)=n$, und die $h_n(i)/n$ sollten in der Nähe von p_i sein: Abweichungen würden ein Indiz dafür sein, dass die Stichprobe von einem anderen Wahrscheinlichkeitsraum gezogen wurde.

Wir nehmen an, dass n „sehr, sehr groß“ ist und dass wir keine Lust haben, das Experiment wirklich durchzuführen; wie könnte man die Ergebnisse einfacher erzeugen? Für jedes i kann doch der Unterschied zwischen p_i und $h_n(i)/n$ aufgrund des zentralen Grenzwertsatzes beschrieben werden:

$$\frac{h_n(i)-np_i}{\sqrt{np_i(1-p_i)}}$$

ist annähernd standard-normalverteilt. Man könnte sich also standard-normalverteilte $\xi_1,\dots,\xi_s$ verschaffen und $h_n(i)$ als $\sqrt{np_i(1-p_i)}\xi_i+np_i$ definieren.

Der Schönheitsfehler: Auf diese Weise kann nicht garantiert werden, dass $h_n(1)+\cdots+h_n(s)=n$ gilt. Das ist nicht zu erwarten, wenn die ξ_i *unabhängig* sind, dann kann $h_n(1)+\cdots+h_n(s)=n$ keine Konstante sein.

Man muss also die ξ_i normalverteilt *mit gewissen Abhängigkeiten* so erzeugen, dass am Ende wirklich die Summe der $h_n(i)$ gleich n ist. Technisch geht das so: Man erzeugt $s-1$ unabhängige $N(0,1)$-Verteilungen $\eta_1,\dots,\eta_{s-1}$ und verwendet sie, um s Standard-Normalverteilungen $\xi_1,\dots,\xi_s$ zu definieren. ξ_i spielt die Rolle von $\big(h_n(i)-np_i\big)/\sqrt{np_i}$. Erstens kann man so die $h_n(i)$ aus den η's errechnen, und zweitens entstehen die η's aus den ξ's durch eine orthogonale Transformation.

Fazit: Die Summe $\xi_1^2+\cdots+\xi_s^2$ ist genauso verteilt wie die $\eta_1^2+\cdots+\eta_{s-1}^2$, sie ist also χ^2_{s-1}-verteilt. Auf diese Weise ergibt sich der folgende

Satz 5.1.1. *(χ^2-Anpassungstest von Pearson, 1900[1)])*
Definiere eine Zufallsvariable D durch

$$D:=n\sum_{i=1}^{s}p_i\left(\frac{h_n(i)/n}{p_i}-1\right)^2.$$

Dann ist D für große n näherungsweise χ^2_{s-1}-verteilt.

Beweis: Der Beweis ist recht aufwändig. Hier sollen die entscheidenden Techniken vorgestellt werden, für Einzelheiten verweise ich auf das Buch von Georgii.

Schritt 1: Was ist die Multinomialverteilung?

[1)] Von K. Pearson, Vater von E. Pearson aus dem Neyman-Pearson-Lemma.

Die *Binomialverteilung* ist wohlbekannt: Es werden n unabhängige Experimente gemacht, die jeweils mit Wahrscheinlichkeit p zum „Erfolg“ führen. Es geht dann um die Wahrscheinlichkeit, dass genau k Erfolge dabei sind.

Bei der *Multinomialverteilung* gibt es auch n Experimente, Es gibt aber „Erfolge“ vom Typ 1 bis s, der i-te Erfolg tritt mit Wahrscheinlichkeit p_i ein. Unter der *Multinomialverteilung* versteht man das Wahrscheinlichkeitsmaß, das die Verteilung der möglichen Ergebnisse misst:
Wie groß ist – für $j_1, \ldots, j_s \in \mathbb{N}_0$ – die Wahrscheinlichkeit, dass genau j_i Erfolge des Typs i dabei sind, $i = 1, \ldots, s$?

Die Antwort bekommt man durch sukzessive Anwendung der Idee, die zur Binomialverteilung führte. Wir betrachten ein Experiment, das genau j_i Erfolge vom Typ i hatte (für alle $i = 1, \ldots, s$). Jedes einzelne hat sicher die Wahrscheinlichkeit $p_1^{j_1} \cdots p_s^{j_s}$. Wieviele gibt es?

Es gibt $\binom{n}{j_1}$ Möglichkeiten, die 1-Erfolge zu wählen, unter den verbleibenden $n - j_1$ Experimenten gibt es $\binom{n-j_1}{j_2}$ Möglichkeiten für die 2-Erfolge. Und so weiter. Am Ende erhält man

$$\frac{n!}{j_1! \cdots j_s!}$$

mögliche Fälle. Die gesuchte Wahrscheinlichkeit ergibt sich damit zu

$$M(j_1, \ldots, j_s; p_1, \ldots, p_s; n) := \frac{n!}{j_1! \cdots j_s!} p_1^{j_1} \cdots p_s^{j_s}.$$

Das ist die Multinomialverteilung.

Schritt 2: Wie hängt sie mit unserem Problem zusammen?
Aufgrund unserer Herleitung sind die Erfolgszahlen $(h_n(1), \ldots, h_n(s))$ gemäß $M(\ldots; p_1, \ldots, p_s; n)$ verteilt. Wir werden daher die Multinomialverteilung studieren müssen, die Behauptung des Satzes wird sich daraus ergeben, dass wir alles (approximativ) auf unabhängige Summen von $N(0,1)$-Verteilungen zurückführen.

Schritt 3: Das Maß P_0 auf der Hyperebene E
Eine wichtige Rolle wird im Folgenden die durch

$$E := \{x \in \mathbb{R}^s \mid \sum_i x_i \sqrt{p_i} = 0\}$$

definierte Hyperebene des $\mathbb{R}^s$ spielen.

Wir stellen uns die $h_n(i)$ $(i = 1, \ldots, s)$ nämlich als Zufallsvariable vor[2)]. Normalisiert man sie durch die Definition

$$h_n^*(i) := \big(h_n(i) - np_i\big)/\sqrt{np_i},$$

[2)] Genauer: Wähle n unabhängige Zufallsvariable $Y_1, \ldots, Y_n$ mit Werten in $\{1, \ldots, s\}$, so dass das Bildmaß jeweils die durch $p_1, \ldots, p_s$ gegebenen Wahrscheinlichkeiten hat. Es ist dann

$$h_n(i) = \sum_j \chi_{\{i\}}(Y_j).$$

so entsteht ein Zufallsvektor

$$\big(h_n^*(1),\dots,h_n^*(s)\big),$$

der offensichtlich in E liegt. Unter P_0 wollen wir das durch diesen Vektor induzierte Bildmaß verstehen.

> Dabei unterdrücken wir das n:
> **Ab jetzt ist n fixiert, es wird angenommen, dass n „sehr groß“ ist.**

Schritt 4: Auch Poissonverteilungen erzeugen die Multinomialverteilung

Einleitend wurde schon bemerkt, dass die Hauptschwierigkeit darin besteht, die betrachteten Verteilungen durch *unabhängige* Normalverteilungen zu beschreiben. Mit unabhängigen Poissonverteilungen geht das wie folgt.

Betrachte für $i = 1,\dots,s$ eine Zufallsvariable S_i, die Poissonverteilt zum Parameter np_i ist; die $S_1,\dots,S_s$ sollen dabei unabhängig sein. Schon jetzt bemerken wir, dass S_i als Summe von n unabhängigen Poissonverteilungen mit Parameter p_i erzeugt werden kann; deswegen ist $S_i^* := (S_i - np_i)/\sqrt{np_i}$ annähernd standard-normalverteilt[3)].

Wir setzen noch $N := \sum_i S_i$. Diese Zufallsvariable ist Poissonverteilt zum Parameter n, und deswegen ist $N^* := (N-n)/\sqrt{n}$ ebenfalls annähernd standard-normalverteilt.

Betrachte nun das Ereignis $\{N = n\}$ (d.h. $\{N^* = 0\}$). Nach Definition der Poissonverteilung hat es die Wahrscheinlichkeit $n^n\mathrm{e}^{-n}/n!$. Entsprechend erhält man: Für $j_1,\dots,j_s \in \mathbb{N}_0$ mit $\sum_i j_i = n$ ist die Wahrscheinlichkeit

$$P(S_i = j_i \text{ für alle } i \mid N = n)$$

gleich

$$\frac{(np_1)^{j_1}\mathrm{e}^{-np_1}}{j_1!}\cdots\frac{(np_s)^{j_s}\mathrm{e}^{-np_s}}{j_s!}\Big/\, n^n\mathrm{e}^{-n}/n! \quad = \quad M(j_1,\dots,j_s;p_1,\dots,p_s;n).$$

So kommen unabhängige Zufallsvariable ins Spiel.

Schritt 5: Bedingte Normalverteilungen

Wir betrachten s unabhängige Standard-Normalverteilungen $\xi_1,\dots,\xi_s$. Dann ist der Vektor $\xi := (\xi_1,\dots,\xi_s)$ gemeinsam normalverteilt auf dem $\mathbb{R}^s$. Wir benötigen gleich das Ergebnis, dass ξ unter der Bedingung $\xi_s = 0$ gemeinsam wie $(\xi_1,\dots,\xi_{s-1})$ auf dem $\mathbb{R}^{s-1}$ normalverteilt ist. Zur Präzisierung dieses plausiblen und richtigen Ergebnisses müssten noch einige Einzelheiten nachgetragen werden, die wir hier überspringen.

Schritt 5: Finale (Hauptbeweis)

Sei O eine orthonormale Matrix, die die Hyperebene $H := \{x_s = 0\}$ des $\mathbb{R}^s$ in E dreht.

[3)] Zur Erinnerung: Die Varianz einer Poissonverteilung zum Parameter λ ist gleich λ.

Wähle dazu eine orthonormale Matrix V, in deren letzter Zeile der Einheitsvektor $\left(\sqrt{p_i}\right)$ steht. Dann vermittelt V durch $x \mapsto Vx$ offensichtlich eine Bijektion von E nach H. Also reicht es, $O := V^{-1}$ zu setzen.

Da O orthonormal ist, sind auch die durch

$$Y := \left(Y_1^*, \ldots, Y_s^*\right)^\top := O\left(S_1^*, \ldots, S_s^*\right)^\top$$

definierten Zufallsvariablen $Y_1^*, \ldots, Y_s^*$ (näherungsweise) gemeinsam standardnormalverteilt und unabhängig. Wegen Schritt 5 ist also Y^* unter der Bedingung $Y_s^* = 0$ (approximativ) verteilt wie $s-1$ unabhängige Standardnormalverteilungen.

Bemerkenswerterweise ist aber $Y_s^* = N^*$, also gilt:
Die Verteilung von $S_1^*, \ldots, S_s^*$ unter der Bedingung $N^* = 0$ – d.i. die Multinomialverteilung – entsteht durch „Drehung“ von $s-1$ unabhängigen Standardnormalverteilungen im $\mathbb{R}^{s-1} \times \{0\}$.

Da schließlich Längen beim Drehen erhalten bleiben, muss die quadrierte Länge von $\left(S_1^*, \ldots, S_s^*\right)$ wie χ^2_{s-1}-verteilt sein. Die $h_n^*(i)$ entstehen aber aus den $h_n(i)$ wie die S_i^* aus den S_i, und daswegen folgt, dass die quadrierte Länge des Vektors $(h_n^*(1), \ldots, h_n^*(i))$ ebenfalls χ^2_{s-1}-verteilt sein muss. Es ist aber

$$\begin{aligned}
\sum \left(h_n^*(i)\right)^2 &= \sum \frac{(h_n(i) - np_i)^2}{np_i} \\
&= n \sum p_i \left(\frac{h_n(i)/n}{p_i} - 1\right)^2 \\
&= D.
\end{aligned}$$

Damit ist alles gezeigt.

Es folgt ein Versuch, die Strategie noch einmal mit anderen Worten zusammenzufassen. Neben den schon eingeführten Definitionen betrachten wir noch die Abbildung $\Phi : \mathbb{R}^s \to \mathbb{R}^s$, die durch

$$(x_1, \ldots, x_s) \mapsto \left(\frac{x_1 - np_1}{\sqrt{np_s}}, \ldots, \frac{x_s - np_s}{\sqrt{np_s}}\right)$$

definiert ist.

- Die Multinomialverteilung ist eine Verteilung auf

  $$F := \{(j_1, \ldots, j_s) \in (\mathbb{N}_0)^s \mid \sum j_i = n\}.$$

  So sind die $(h_n(1), \ldots, h_n(s))$ verteilt.
- $(S_1, \ldots, S_s)$ ist in $(\mathbb{N}_0)^s$ verteilt, auf F ergibt sich die Multinomialverteilung.
- Nun wird Φ auf die $(S_1, \ldots, S_s)$ angewandt. Die Bildverteilung ist näherungsweise ein Produkt von s unabhängigen Standardnormalverteilungen (zentraler Grenzwertsatz). Deswegen sollte die bedingte Verteilung auf *jeder* Hyperbene durch 0 verteilt sein wie $s-1$ unabhängige Standardnormalverteilungen. Unter Φ wird aber F auf E abgebildet, und wir erhalten:

- Die $(h_1^*, \ldots, h_s^*)$ in E sind verteilt wie $s-1$ unabhängige Standardnormalverteilungen. Deswegen ist die quadrierte Länge dieses Vektors (das ist das D des Satzes) näherungsweise χ^2_{s-1}-verteilt.

□

Bemerkungen und Beispiele:

1. Die ungewöhnliche Normierung scheint auf den ersten Blick etwas mysteriös. Das Ergebnis soll deswegen im Spezialfall $s = 2$ noch einmal durch eine direkte Rechnung verifiziert werden.

Gegeben seien also eine Wahrscheinlichkeit p, es werde n-mal ein Bernoulliexperiment durchgeführt. Mit h bezeichnen wir die Anzahl der Erfolge.

Dann ist $h^* = (h-np)/\sqrt{np(1-p)}$ nach dem zentralen Grenzwertsatz näherungsweise $N(0,1)$-verteilt. Nun ist nur zu beachten, dass im vorliegenden Fall

$$(h^*)^2 = n\Big(p\Big(\frac{h/n}{p} - 1\Big)^2 + (1-p)\Big(\frac{(n-h)/n}{1-p} - 1\Big)^2\Big)$$

gilt, und die rechte Seite ist der Testwert D des Satzes.

2. In den Lehrbüchern gibt es ein beliebtes Beispiel: Das Mendelsche Kreuzungsexperiment. Es stellt sich nämlich heraus, dass der χ^2-Test, angewandt auf die Mendelschen Zahlen, eigentlich viel zu perfekt ausfällt. Zwischen den Zeilen steht dann immer der Verdacht, dass Mendel die Zahlen vielleicht ein wenig manipuliert hat: corriger la fortune ...

3. Die Ausgangsfrage „Wie fälscht man Protokolle richtig?“ ist jetzt auch beantwortet. Man verschaffe sich

- $s-1$ unabhängige $N(0,1)$-Verteilungen $\xi_1, \ldots, \xi_{s-1}$.
- Eine orthonormale Matrix O, die in der letzten Spalte den Vektor $(\sqrt{p_1}, \ldots, \sqrt{p_s})^\top$ hat; die Einträge sollen a_{ij} heißen.

Definiert man dann $A_i := \sum_{j=1}^{s-1} a_{ij}\xi_j$ und anschließend

$$h_i := \sqrt{np_i}A_i + np_i$$

für $i = 1, \ldots, s$, so haben die h_i die richtige Verteilung; man muss nur noch auf ganzzahlige Werte auf- bzw. abrunden.

5.2 Der χ^2-Unabhängigkeitstest

Überraschenderweise kann man mit ähnlichen Methoden auch Unabhängigkeit testen. Die Ausgangssituation ist die folgende.

Gegeben sind zwei endliche Mengen $I = \{1, \ldots, k\}$ und $J = \{1, \ldots, l\}$ sowie Wahrscheinlichkeiten p_{ij} auf $I \times J$. Die *Marginalverteilungen* sind dann die $p_i := \sum_j p_{ij}$ und die $q_j := \sum_i p_{ij}$. Man kann sich dann fragen, ob die Komponenten i, j *unabhängig* sind, ob also das Maß auf $I \times J$ das Produktmaß der Marginalverteilungen ist.

Beispiele aus dem Leben sind schnell gefunden: Sind die Eigenschaften „Raucher“ und „Geschlecht“ unabhängig (hier ist $k = l = 2$)? Liegt Unabhängigkeit vor, wenn man das Gehalt und die Uni vergleicht, an der der Abschluss gemacht wurde?

Das einzige, was man messen kann, sind die relativen Häufigkeiten in einer Stichprobe. Man nimmt also eine Stichprobe vom Umfang n und zählt, wie sich die Merkmale verteilen (wieviele weibliche Raucher usw.). Wenn es h_{ij} Ergebnisse des Typs (i, j) gibt, so ist also

- erstens $\sum_{ij} h_{ij} = n$;
- h_{ij}/n eine Schätzung für p_{ij};
- Unabhängigkeit genau dann gegeben, wenn h_{ij}/n in der Nähe des Produkts aus $(\sum_j h_{ij})/n$ und $\sum_i h_{ij}/n$ ist.

Die Idee des χ^2-Tests auf Unabhängigkeit besteht nun darin, aus den h_{ij} eine Testgröße zu berechnen, die im Fall der Unabhängigkeit nach einer bekannten Verteilung verteilt ist. Das ist der Inhalt des folgenden Satzes:

Satz 5.2.1. *Mit den vorstehenden Bezeichnungen gilt: Setzt man* $h_i^1 := \sum_j h_{ij}$ *und* $h_j^2 := \sum_i h_{ij}$ *sowie*

$$T := \sum_{ij} \frac{\left(h_{ij} - h_i^1 h_j^2/n\right)^2}{h_i^1 h_j^2/n},$$

so ist T *näherungsweise* $\chi^2_{(k-1)(l-1)}$*-verteilt.*

Bemerkungen:

1. Der Beweis ist ähnlich wie der zum χ^2-Anpasssungstest: Man überlegt sich, wie man eine Stichprobe mit $(k-1)(l-1)$ unabhängigen Standardnormalverteilungen erzeugen kann und rechnet dann rückwärts aus den h_{ij} eine Zahl aus, die der Quadratsumme dieser Verteilungen entspricht: vgl. Georgii, Abschnitt 11.3.

2. Bei einer typischen Anwendung gibt man ein α vor, macht den Test und lehnt dann die Hypothese der Unabhängigkeit ab, wenn T zu groß ausfällt (wenn also $T > r$ ist, wobei eine $\chi^2_{(k-1)(l-1)}$-Verteilung mit Wahrscheinlichkeit $1 - \alpha$ in $[\,0, r\,]$ liegt).

5.3 Rangtests

Wie kann man testen, ob zwei Stichproben aus der gleichen Verteilung gezogen wurden? Um entsprechende Tests zu motivieren, betrachten wir vorbereitend die folgende Situation: Es treten bei einem Mathematikwettbewerb zwei Teams an, Team A und Team B. Wenn dann Team A „deutlich besser“ ist, findet es

niemand überraschend, dass die Platzierung (von schlecht nach gut) etwa so aussieht:

$$B, A, B, B, B, A, B, B, B, B, B, A, A, A, A, A, A, A.$$

Bei „ausgeglichenen“ Teams würde man eher Ergebnise des Typs

$$A, B, B, A, B, A, A, B, B, B, A, B, A, A, B, A, A, B$$

erwarten. Kurz: Eine ausgeglichene Rangfolge ist ein Indiz für ein vergleichbares Potenzial. Diese Idee möchte man sich zum Testen zunutze machen.

Als *erste Vorbereitung* soll mit wahrscheinlichkeitstheoretischen Methoden präzisiert werden, was „besser“ heißt:

Definition 5.3.1. *Seien P und Q zwei Wahrscheinlichkeitsmaße auf $\mathbb{R}$. Wir sagen, dass P* nicht besser *als Q ist, wenn*

$$P([\,c, +\infty\,[) \leq Q([\,c, +\infty\,[)$$

für jedes $c \in \mathbb{R}$ gilt; in diesem Fall schreiben wir $P \prec Q$.

Beispiele: Für zwei Normalverteilungen mit der gleichen Streuung gilt $P \prec Q$ genau dann, wenn der Erwartungswert von P kleiner oder gleich dem Erwartungswert von Q ist. Auch führen bei Binomialverteilungen größere p zu größeren Verteilungen im Sinne von $\prec$.

Das Ziel wird also darin bestehen, eine Testgröße einzuführen, mit der man die Nullhypothese $P = Q$ gegen die Alternativhypothese $P \neq Q$ testen kann. Es folgt die für diesen Zweck angemessene

Definition 5.3.2. *Gegeben seien k Zahlen $x_1, \ldots, x_k$, die aus einer Verteilung P gezogen wurden sowie l Zahlen $x_{k+1}, \ldots, x_{k+l}$, die gemäß der Verteilung Q erzeugt wurden. Wir nehmen an, dass die $x_1, \ldots, x_{k+l}$ paarweise verschieden sind*[4)].

Und nun wird gezählt: Die Zahl U entsteht daraus, dass man für jedes x_i, $i = 1, \ldots, k$, zählt, wieviele x_j mit $j = k+1, \ldots, k+l$ vor x_i liegen:

$$U := \sum_{i=1}^{k} \sum_{j=k+1}^{k+l} \chi_{x_j < x_i}.$$

Beispiel:

Es sei etwa $k = l = 3$, die x_i seien gleich $1, 4, 7, 2, 3, 10$. Dann ist $U = 0+2+2 = 4$.

Man beachte, dass U stets zwischen 0 und kl liegt. (Im ersten Fall sind die $x_1, \ldots, x_k$ die kleinsten Elemente der Stichprobe, im zweiten Fall die größten.)

[4)] Das ist mit Wahrscheinlichkeit 1 zu erwarten, wenn stetige Dichten vorliegen.

Es gibt eine andere Möglichkeit, den Begriff „besser“ zu quantifizieren. Dazu stelle man sich die geordneten Ergebnisse der aus k bzw. l Mitgliedern bestehenden Teams wie im ersten Absatz als Rangfolge vor. Dann werden die Ränge aus Team 1 (bzw. Team 2) addiert, das führt zu den Zahlen W_1 und W_2. Offensichtlich ist $W_1 + W_2 = 1 + \cdots + n = n(n+1)/2$, deswegen reicht es, zur Beschreibung der Situation die Zahl $W := W_1$ anzugeben.

Auch der Zusammenhang zu U kann leicht beschrieben werden:

$$W = U + k(k+1)/2.$$

(Begründung: Sei o.E. die erste Gruppe schon angeordnet. Für das i-te Mitglied dieser Gruppe gilt: Sein Rang ist i plus die Anzahl aus der zweiten Gruppe, die davor liegen; das „i“ kommt daher, dass so viele vom eigenen Team bei der Rangfolge zu berücksichtigen sind. Summiert man über alle i auf, so entsteht links W und rechts $1 + \cdots + k + U$.

Ein Beispiel: Für die vorstehend angegebene Stichprobe ist $W_1 = 10$ und $W_2 = 11$. Erwartungsgemäß ist erstens $W_1 + W_2 = 6 \cdot 7/2$ sowie $W_1 = U + 3 \cdot 4/2$.

Bemerkenswerterweise kann man die Verteilung von U berechnen:

Satz 5.3.3. *Die $x_1, \ldots, x_k, x_{k+1}, \ldots, x_{k+l}$ sei eine Stichprobe, die aus einer Verteilung mit einer stetigen Dichte gezogen wurde. $0 \le m \le kl$. Die Wahrscheinlichkeit, dass $U = m$ gilt, ist durch*

$$N(m; k, l) \;/\; \binom{n}{k}$$

gegeben; dabei ist $n = k + l$ und $N(m; k, l)$ die Anzahl der Möglichkeiten, die Zahl m als

$$m = m_1 + \cdots + m_k \text{ mit } 0 \le m_1 \le \cdots \le m_k \le l$$

zu schreiben.

Beweis: Da die Verteilung eine stetige Dichte hat, stimmen zwei x_i der Stichprobe nur mit Wahrscheinlichkeit 0 überein, wir dürfen also annehmen, dass sie alle verschieden sind; dadurch ist die Rangfolge eindeutig. Aufgrund der Unabhängigkeit ist $x_i < x_j$ gleichwahrscheinlich zu $x_j < x_i$, und deswegen ist die Verteilung der Rangfolgen gleichverteilt auf der Menge der Permutationen der Menge $\{1, \ldots, n\}$.

Nun werden die $x_1, \ldots, x_k$ zufällig ausgewählt. Wir müssen berechnen, wieviele Möglichkeiten es dafür gibt und welche davon zu $U = m$ führen. Die gesuchte Wahrscheinlichkeit ist dann der Quotient („günstige Fälle durch mögliche Fälle“).

Die erste Anzahl (der Nenner) ist leicht zu bestimmen:

$$n(n-1)\cdots(n-k+1).$$

Für die Anzahl im Zähler fixieren wir die $x_1, \ldots, x_k$. Wir konzentrieren uns zunächst auf den Fall $x_1 < \cdots < x_k$ und prüfen dafür nach, ob $U = m$ herauskommt. Wenn ja, müssen wir diese Zahl noch mit $k!$ multiplizieren, denn auf so viele Weise können die $x_1, \ldots, x_k$ permutiert werden.

Angenommen also, das geordnete k-Tupel führt zu $U = m$. Vor x_1 liegen dann m_1 aus der Menge $\Delta = \{x_{k+1}, \ldots, x_n\}$, vor x_2 liegen m_2 aus Δ usw. bis x_k. Dann ist erstens $m_1 \leq \cdots \leq m_k$, zweitens gilt $m_1, \ldots, m_k \in \{0, 1, \ldots, l\}$ und drittens schließlich ist $m_1 + \cdots + m_k = m$.

Zusammen: Im geordneten Fall gibt es $N(m; k, l)$ Möglichkeiten, die zu $U = m$ führen, insgesamt erhält man also $k!N(m; k, l)$. Für die Wahrscheinlichkeit erhält man also

$$\frac{k!N(m;k,l)}{n(n-1)\cdots(n-k+1)} = N(m;k,l) \;/\; \binom{n}{k}.$$

□

Die U-Statistik hängt also nicht von der speziellen Verteilung ab und kann Tafeln entnommen werden. Man spricht von einem *Mann-Whitney-Test* oder auch einem *Wilcoxon-Test*, wenn aufgrund dieser Statistik die Hypothese $P = Q$ zu einem vorgelegten Niveau α behandelt werden soll.

Beispiel: Es sollen zwei Schlafmittel verglichen werden: Ist das eine besser? Die Schlafdauer wird in jeweils 10 Versuchen gemessen, es ergibt sich $U = 25$. Nach dem vorstehenden Satz kann die Hypothese $P = Q$ zum Niveau $\alpha = 0.025$ nicht abgelehnt werden.

5.4 Der Kolmogoroff-Smirnoff-Test

Mit dem χ^2-Anpassungstest konnte man die Hypothese behandeln, ob eine Stichprobe aus einer vorgegebenen Verteilung auf einem *endlichen* Wahrscheinlichkeitsraum gezogen wurde. In diese Abschnitt geht es um die analoge Frage für den Fall von Wahrscheinlichkeitsverteilungen mit Dichten.

Gegeben sei also ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$, das durch eine stetige Dichte f definiert ist. Dann ist die Verteilungsfunktion F eine differenzierbare Funktion von $\mathbb{R}$ nach $\mathbb{R}$. Es soll die Hypothese betrachtet werden, dass eine Stichprobe $x_1, \ldots, x_n$ von genau diesem Wahrscheinlichkeitsraum gezogen wurde.

Die Idee beim Kolmogoroff-Smirnoff-Test besteht darin, die Stichprobe zur Definition der *empirischen Häufigkeitsverteilung* zu verwenden und die Ablehnung der Hypothese davon abhängig zu machen, wie weit diese von F entfernt

ist. Genauer: Wir definieren

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \chi_{[\,x_i,+\infty\,[}(x)$$

und betrachten dann

$$D_n := \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|.$$

Falls die x_i wirklich gemäß F erzeugt wurden, sollte D_n in der Regel klein sein. Wenn es gelänge, die Verteilung von D_n zu berechnen, könnte man das verwenden, um die Hypothese zu vorgegebenen Niveaus abzulehnen.

Überraschenderweise hängt die Verteilung von D_n nicht von F ab. Deswegen muss man sie nur ein einziges Mal berechnen: *Das* ist die *Kolmogoroff-Smirnoff-Verteilung*. (Ein weiteres – weniger wichtiges – Problem besteht dann darin, eine geschlossene Formel für die Verteilung anzugeben; eine Lösung wäre entbehrlich, da ja auch andere Möglichkeiten existieren, Tafeln für den praktischen Gebrauch zu berechnen.)

$\boxed{D_n \text{ ist unabhängig von } F}$

Es ist für die Rechnungen unbequem, dass in der Definition von D_n ein Betrag auftritt. Deswegen machen wir einen kleinen Umweg. Wir definieren:

$$D_n^+ := \sup_{x\in\mathbb{R}} \big(F_n(x) - F(x)\big),$$

$$D_n^- := \sup_{x\in\mathbb{R}} \big(F(x) - F_n(x)\big) = - \inf_{x\in\mathbb{R}} \big(F_n(x) - F(x).$$

Es ist klar, dass dann gilt:

$$D_n = \max\{D_n^+, D_n^-\}.$$

Deswegen reicht es, D_n^+ detaillierter zu untersuchen (die Rechnungen für D_n^- ergeben sich dann durch Übergang von „sup“ zu „inf“.)

Entscheidend sind die folgenden *Beobachtungen*:

- Angenommen, eine Zufallsvariable X ist so verteilt, dass P_X die Verteilungsfunktion F hat. Dann ist $F(X)$ gleichverteilt in $[\,0,1\,]$.

  Begründung: Sei $[\,a,b\,]$ ein Teilintervall von $[\,0,1\,]$. Dann sind offensichtlich äquivalent:

  - $F(X)$ liegt in $[\,a,b\,]$.
  - X liegt in $[\,\alpha,\beta\,]$; dabei ist $\alpha := F^{-1}(a)$ und $\beta := F^{-1}(b)$.

  Deswegen ist die Wahrscheinlichkeit für das eben beschriebene Ereignis gleich

  $$F\big(F^{-1}(b)\big) - F\big(F^{-1}(a)\big),$$

  also gleich $b - a$. Das beweist die Behauptung.

- Zur Berechnung von D_n^+ spielt es keine Rolle, wie die Stichprobe angeordnet ist: O.B.d.A. gilt $x_1 < \cdots < x_n$.
- Für x zwischen x_i und x_{i+1} ist $F_n(x) = i/n$, und deswegen gilt

$$\begin{aligned} D_n^+ &= \max_{0\le i\le n} \sup_{x_i\le x\le x_{i+1}} \Big(\frac{i}{n} - F(x)\Big) \\ &= \max_{0\le i\le n} \Big(\frac{i}{n} - \inf_{x_i\le x\le x_{i+1}} F(x)\Big) \\ &= \max_{0\le i\le n} \Big(\frac{i}{n} - F(x_i)\Big) \end{aligned}$$

Zusammen: Um D_n^+ zu simulieren, betrachte man n gleichverteilte, geordnete Zahlen $y_1, \ldots, y_n$ in $[\,0, 1\,]$ und bestimme dann

$$\max_{0\le i\le n} \Big(\frac{i}{n} - y_i\Big).$$

Man beachte, dass das *nicht mehr von* F abhängt. (Für D_n^- muss man in der vorstehenden Rechnung „max“ durch „min“ ersetzen.)

Die explizite Form von D_n

Wir geben nur das Ergebnis an:

Satz 5.4.1. *Für jedes t ist*

$$P(\sqrt{n}D_n \le t) = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} \mathrm{e}^{-2j^2t^2}.$$