Statistische Methoden

Dr. C.J. Luchsinger

3 Grundlagen der Statistik

Literatur Kapitel 3: Lindgren: Kapitel 7

3.1 Überblick

Die Statistik ist ein ausuferndes Wissensgebiet. Man könnte problemlos 20 Semester lang Vorlesungen über Statistik besuchen - so viele Probleme und Methoden gibt es (ein Blick auf www.math-jobs.com/journals.html illustriert, in wie vielen Wissensgebieten die Statistik ihren Siegeszug angetreten hat)! Im Gegensatz zur Wahrscheinlichkeitstheorie, wo wir mit P (Kapitel 1), X (Kapitel 2), E (Kapitel 3), $n \to \infty$ (Kapitel 5) den klassischen Aufbau beschritten haben, gibt es in der Statistik keinen kanonischen Aufbau. Zumindest die 3 ersten Themen, welche wir behandeln (Testen, Schätzen und Regression) gehören aber ganz sicher zu jeder statistischen Ausbildung. Wir treffen also ganz sicher keine falsche Auswahl, wenn wir diese Themen behandeln.

Im Buch "Statistische Datenanalyse" von Stahel wird in den drei ersten Kapiteln den Fragen "Was ist statistische Datenanalyse, Ziele" und "Beschreibende Statistik" nachgegangen. Ich empfehle, diese Kapitel zu lesen, entweder jetzt oder im Sommer. Man kann aber auch ohne diese Kapitel problemlos diese Vorlesung besuchen. Vieles, was dort besprochen wird, ist zwar sehr wichtig, jedoch nicht für eine Vorlesung in mathematischer Statistik geeignet. Der "Cartoon Guide to Statistics" von Gonick und Smith geht in den ersten beiden Kapiteln ebenfalls diesen Fragen nach.

Ohne Anspruch auf Vollständigkeit, stichwortartig einige Punkte aus obiger Literatur:

* Statistik: Daten $\Rightarrow$ Schlüsse ziehen

* Zwei Arten von Statistik: exploratorisch und konfirmatorisch (z.B. bei Medikamentenentwicklung und -Zulassung): In der *exploratorischen Statistik* (engl. to explore

= erforschen), wird einE StatistikerIn bei der Medikamentenentwicklung beigezogen, um aus Datenmaterial zu Krankheitsverläufen und Wirkungsweisen von Substanzen Hinweise für PharmazeutInnen, ChemikerInnen, BiologInnen und MedizinerInnen zu geben, wo eventuell weiter gesucht werden sollte. StatistikerInnen sind hierbei ganz klar nur Teil eines Teams und können in keiner Art und Weise die anderen WissenschaftlerInnen ersetzen oder in den Schatten stellen. Sie liefern nur eine Teilsicht des Problems - Intuition der anderen WissenschaftlerInnen können viel bedeutender sein. Bei der Wahl von statistischen Methoden ist man hier sehr frei - gut ist, was Erfolg bringt (gilt v.a. auch in der Finanzindustrie). Hat man dann einen Kandidaten für ein erfolgreiches Medikament, muss man zusammen mit den Zulassungsbehörden (in den USA die FDA) testen, ob es die gewünschten Eigenschaften hat und ob die Nebenwirkungen vertretbar sind. Dies geschieht in der *konfirmatorischen Statistik* (engl. to confirm = bestätigen; der Pharmakonzern behauptet dies und das, wir werden mit der Testserie versuchen, diese Resultate zu bestätigen). Dabei wird man in verschiedenen Phasen I-IV nach allfälligen Tierversuchen das Medikament unter anfänglich extrem aufmerksamer Beobachtung an immer mehr Menschen testen.

* Klassische Statistik / Frequenzialisten / Bayesianer / Persönliche Wahrscheinlichkeiten: Es gibt verschiedene Philosophien, mit denen Statistik betrieben werden kann. Wichtig und erfolgreich in Life Sciences und Finanzindustrie eingesetzt ist Bayesianische Statistik. Ein Crash-Course in Bayesianische Statistik bringt aber nichts! Nur ein intensives Studium dieser Methoden, vorausgesetzt sie sprechen einen an, ermöglicht einen erfolgreichen Einsatz dieser Methoden. Sie werden in meinen drei Vorlesungen deshalb *nicht* behandelt. Wir machen hier klassische Statistik.

* Parametrische und nicht-parametrische Statistik: Wenn wir mit Sätzen wie "Sei X eine $\mathcal{N}(\mu, \sigma^2)$-Zufallsgrösse" anfangen, so betreiben wir parametrische Statistik (μ, σ^2 sind Parameter!). Es gibt aber auch nicht-parametrische Methoden (viele Tests zum Beispiel). Die nicht-parametrischen Methoden haben grosse Vorteile (keine einengende Wahl einer expliziten Verteilung). Wir werden in dieser Vorlesung einige nichtparametrische Methoden besprechen (diese Vorlesung hat aber nicht diesen Schwerpunkt).

3.2 Stichproben und empirische Verteilung

* Bis jetzt haben wir immer Wahrscheinlichkeits*theorie* gemacht.
* Die Wahrscheinlichkeitstheorie zeichnet sich dadurch aus, dass wir immer sicher wissen, wie das Modell ist (z.B. "Sei X eine $\mathcal{N}(0,1)$-Zufallsgrösse."). Wir müssen uns in der Theorie *nie* Gedanken machen, ob dieses Modell überhaupt "stimmt".
* Ich setze ab hier voraus, dass Sie, abgesehen von den jeweils angegebenen Voraussetzungen, keine weiteren Kenntnisse oder sonstigen übersinnliche Fähigkeiten haben, welche Ihnen helfen, statistische Schlüsse zu ziehen.
* In der Statistik gilt folgendes: **Wir haben nur die Daten** $(x_1, x_2, x_3, \ldots, x_n)$**!!!** und wissen nicht, aus welcher Verteilung die eigentlich stammen. Diese Daten können Würfelaugen sein bei n Würfen, Blutdruckmessungen bei verschiedenen Personen oder bei der gleichen Person zu verschiedenen Zeitpunkten, Aktienkurse etc.

So ist die Lage. Was jetzt folgt, ist nicht zwingend als Analysemethode. Es ist ein Vorschlag unter vielen. Es gibt aber gute Gründe, am Anfang und unter anderem folgende Untersuchungen zu machen (mehr dazu eben in der Literatur, welche in 3.1 angegeben ist):

* Sortieren der Daten (Daten z.B. im Vektor d abgelegt) nach der Grösse, **R**: sort(d). Wir erhalten damit die sogenannte Ordnungsstatistik $(x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)})$. Wir haben damit aber immer noch n Datenpunkte - im Gegensatz zu gleich nachfolgenden Zusammenfassungen der Daten:
* grösster und kleinster Wert, beide zusammen, Median, **R**: max(d), min(d), range(d), median(d)
* Histogramm, **R**: hist(d)
* arithmetisches Mittel, Stichproben-Varianz, **R**: mean(d), var(d)

Beim letzten Punkt (arithmetisches Mittel $\overline{x}$, Stichproben-Varianz s^2) wollen wir ein bisschen verweilen. Per Definitionem gilt hierfür:

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

und

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

Es gilt übrigens (warum? entweder mühsam nachrechnen oder...):

$$s^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2.$$

Bei der Definition von $\overline{x}$ wird offensichtlich, dass wir hier eigentlich einen Erwartungswert berechnet haben, indem wir jedem Datenpunkt x_i das Gewicht (die Wahrscheinlichkeit)

$$\frac{1}{n}$$

gegeben haben. Dasselbe haben wir bei der Stichproben-Varianz s^2 auch gemacht (in R wird var(d) leicht anders berechnet, man teilt dort durch $(n-1)$ - mehr dazu in Kapitel 5).

Wir sind von der Theorie zu den Daten gegangen und wollen jetzt unser Augenmerk auf ein Zwischending lenken: Wir können doch in einem Statistikpaket wie R einfach eine 100-er Stichprobe von Daten nehmen, von denen wir wissen, dass sie aus einer $\mathcal{N}(0,1)$-Zufallsgrösse stammen: rnorm(100). Wir setzen hier voraus, dass R *für unsere einfachen Berechnungen* einen guten Zufallsgenerator hat. Der gewaltige Vorteil von Simulationen für die Statistik ist der, dass wir wissen, dass z.B. der Erwartungswert 0 war. mean(d) wird aber gerade bei einer $\mathcal{N}(0,1)$-Zufallsgrösse niemals genau 0 sein. Wir können aber so feststellen, wie gut ein Schätzer wie das arithmetische Mittel für die Schätzung des Erwartungswertes ist. Mehr dazu in Kapitel 5. Man mache sich klar, dass wir mit realen Daten nicht wissen, wie die Verteilung eigentlich aussieht und was der Erwartungswert ist;

wenn wir mean(d) eingeben, kommt einfach eine reelle Zahl heraus... . So what?

Jargon:

* In der Statistik mit realen Daten aus der Welt sprechen wir von *Daten oder Stichproben*; bei Statistikpaketen oder wenn wir die Theorie anhand eines Beispiels veranschaulichen wollen, sprechen wir von *Realisationen*: "Sei $x_1, x_2, \ldots, x_n$ eine Realisation vom Umfang n aus einer $\mathcal{N}(0,1)$-Zufallsgrösse".

* Daten werden immer mit kleinen Buchstaben angegeben. Meist werden wir die dazugehörende Zufallsgrösse (aus der die Realisation stammt oder wir gehen zumindest mal davon aus) mit den dazugehörenden Grossbuchstaben bezeichnen: X_i und x_i.

* Wenn nicht anders vereinbart, werden Stichproben/Realisationen vom Umfang n so behandelt, dass man jedem Datenpunkt die Wahrscheinlichkeit $1/n$ zuweist und davon ausgeht, dass die n Daten unabhängig voneinander generiert wurden. Nochmals: Dies wird (mit gutem Grund) vereinbart und ist keineswegs zwingend! Wir haben dann also:

$$(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) = (x_1, x_2, \ldots, x_n),$$

wenn X_i z.B. die Anzahl Augen beim i-ten Wurf ist und die Welt im Zustand ω ist und die X_i's voneinander unabhängig sind.

Wir wollen den Aspekt mit den Gewichten $(1/n)$ noch kurz illustrieren. Wenn man eine $\mathcal{N}(0,1)$-Zufallsgrösse hat, so kommen Werte um 0 häufiger vor, als Werte um 2, 3 oder gar 4. Aber alle Werte der Realisation werden mit $1/n$ gewichtet. Ist das sinnvoll?

Wenn wir eine Stichprobe $(x_1, x_2, \ldots, x_n)$ vom Umfang n haben, so geben wir jedem Punkt x_i die Wahrscheinlichkeit $1/n$ (uniforme Verteilung). Wir nennen dies die Empirische Verteilung. Die empirische Verteilungsfunktion $F_n(x)$ definiert man dann durch

$$F_n(x) := \frac{1}{n}|\{i|x_i \leq x\}| = \frac{1}{n}\sum_{i=1}^{n} I[x_i \leq x].$$

Stammt dann die Stichprobe aus einer Verteilung mit Verteilungsfunktion F_X, so gilt (wenn wir jetzt $F_n(x) := \frac{1}{n}|\{i|X_i \leq x\}|$ als Zufallsgrösse *vor* der Realisierung auffassen):

$$\lim_{n\to\infty} F_n(x) = F_X(x) \quad \forall\, x \in \mathbb{R}.$$

Beweis (zuerst nur Konvergenz in Wahrscheinlichkeit):

□

Bevor wir auf der kommenden Seite den Satz von Glivenko-Cantelli formulieren und beweisen, besprechen wir noch die "fast sichere (fs) Konvergenz" von Zufallsgrössen. Ein Teil der StudentInnen hat dieses Konzept bereits in der Vorlesung "Angewandte Stochastik" (dort nur ein Crash-Kurs) oder in der Vorlesung "Wahrscheinlichkeitstheorie" (ausführlich) kennengelernt. So viel zusammengefasst: bei Theorem 1.24 haben wir bereits fast sichere Konvergenz (deshalb nennt man es auch das Strong Law of Large Numbers SLLN) und in der Folge haben wir auch oben eine fast sichere Konvergenz!

Satz 3.1 [Satz von Glivenko-Cantelli] *Sei $X_i, i \geq 1$ eine iid Folge von Zufallsgrössen mit Verteilungsfunktion $F(x)$. Sei*

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} I[X_i \leq x];$$

($F_n(x) := \frac{1}{n} \sum_{i=1}^{n} I[x_i \leq x]$ ist damit die empirische Verteilungsfunktion einer Stichprobe $x_1, \ldots, x_n$ - wir bezeichnen beide mit F_n). Dann gilt fs:

$$\sup_x |F_n(x) - F(x)| \to 0$$

für $n \to \infty$.

Beweis von Satz 3.1 Für eine reelle Zahl x_1 gilt für $n \to \infty$ fast sicher, dass

$$F_n(x_1) \to F(x_1) \tag{I}$$

(Beweis eine Seite weiter vorne). Es gibt also eine Nullmenge $N_{(1,1)}$, sodass $\forall \omega \notin N_{(1,1)}$ obige Konvergenz (I) gilt. Weiter haben wir für $n \to \infty$ fs

$$F_n(x_1-) := \frac{1}{n} \sum_{i=1}^{n} I[X_i < x_1] \to F(x_1-); \tag{II}$$

wobei wir definieren $F(x_1-) := P[X < x_1]$. Auch hier gibt es demnach eine Nullmenge $N_{(1,2)}$, sodass $\forall \omega \notin N_{(1,2)}$ obige Konvergenz (II) gilt. Abzählbare Vereinigungen von Nullmengen bleiben Nullmengen. Wenn wir also nur endlich viele solche Zahlen $(x_i)_{i=1}^{l}$ (geschickt) auswählen (später), so gilt

$$P[\{\omega | F_n(x_i, \omega) \to F(x_i, \omega) \wedge F_n(x_i-, \omega) \to F(x_i-, \omega) \ \ \forall (x_i)_{i=1}^{l}\}] = 1.$$

Die Ausnahme-Nullmenge bezeichnen wir mit

$$N := \cup_{j=1}^{2} \cup_{i=1}^{l} N_{(i,j)}.$$

Zu gegebenem $\epsilon > 0$ wählen wir eine endliche (!) Folge von Punkten

$$-\infty = x_0, x_1, \ldots, x_l, x_{l+1} = \infty,$$

sodass $F(x_i-) - F(x_{i-1}) < \epsilon/2$, siehe nachfolgendes Bild:

Es existert ein $n_0 := n_0(\omega)$, sodass für alle $n \geq n_0$ gilt:

$$\max_{1 \leq i \leq l+1} \{\max\{|F_n(x_i) - F(x_i)|, |F_n(x_i-) - F(x_i-)|\}\} \leq \epsilon/2.$$

Weiter gelten ausserhalb der Nullmenge N für i so, dass $x \in [x_{i-1}, x_i)$:

$$F_n(x) \geq F_n(x_{i-1}) \geq F(x_{i-1}) - \epsilon/2,$$

sowie

$$F_n(x) \leq F_n(x_i-) \leq F(x_i-) + \epsilon/2$$

und

$$F(x_i-) \leq F(x) + \epsilon/2,$$

sowie

$$F(x_{i-1}) \geq F(x) - \epsilon/2.$$

Damit haben wir für alle $n \geq n_0$ und für alle $x \in \mathbb{R}$, dass

$$|F_n(x) - F(x)| \leq \epsilon/2 + \epsilon/2 = \epsilon.$$

□

3.3 Mathematische Formalisierung

Weil wir hier an einem Mathematik-Institut sind, werden wir klar angeben, in welchen mathematischen Gebilden wir arbeiten. In nachfolgenden Kapiteln verzichten wir dann aber darauf. Wir werden die nachfolgende mathematische Formalisierung von statistischen Problemstellungen immer anhand von 2 Fragestellungen illustrieren: Testen von Hypothesen und Schätzen von Parametern, und zwar im Fall von $\mathcal{N}(\mu, \sigma^2)$-Zufallsgrössen und Stichproben hieraus.

3.3.1 Aktionsraum

Wir haben also eine Stichprobe $(x_1, \ldots, x_n)$, von der wir wissen, dass die $x_i, 1 \leq i \leq n$, unabhängige Realisationen aus einer $\mathcal{N}(\mu, \sigma^2)$-Zufallsgrösse sind. Wenn wir ein *Testproblem* haben, so wollen wir z.B. wissen, ob $\mathcal{H}_0 : \mu = 0$ oder $\mathcal{H}_1 : \mu = 1$ gilt. Der **Aktionsraum** $\mathcal{A}$ beschreibt die Menge der möglichen Aktionen und ist dann gleich

$$\mathcal{A} := \{0, 1\} \quad \text{oder} \quad \mathcal{A} := \{\mathcal{H}_0, \mathcal{H}_1\}.$$

Wenn wir eine Antwort auf die Frage "$\mu = 7$?" geben müssen, so ist der Aktionsraum

$$\mathcal{A} := \{\text{Ja}, \text{Nein}\}.$$

Bei einem *Schätzproblem* (schätzen Sie μ) nehmen wir *in diesem Beispiel* $\mathcal{A} := \mathbb{R}$, weil μ eine beliebige reelle Zahl sein kann.

3.3.2 Entscheidungsfunktion

In einem zweiten Schritt werden wir **Entscheidungsfunktionen** einführen. Wir bemerken vorgängig, dass die Stichprobe als Ganzes im Fall der Normalverteilung ein Element aus dem $\mathbb{R}^n$ ist. Die Menge der Entscheidungsfunktionen $\mathcal{D}$ (engl. decision) ist dann die Menge aller (messbaren) Funktionen:

$$\mathbb{R}^n \to \mathcal{A}.$$

Wenn wir die Stichprobe $(x_1, \ldots, x_n)$ haben, dann wird mit Entscheidungsfunktion d die Aktion $d(x_1, \ldots, x_n) \in \mathcal{A}$ gewählt: z.B. "mit diesen Daten lehne ich $\mathcal{H}_0$ ab" bzw. "schätze $\mu = 0.3544$". In Schätzproblemen ist dann z.B. $d_1(x_1, \ldots, x_n) = \overline{x}$ oder $d_2(x_1, \ldots, x_n) = \text{median}(x_1, \ldots, x_n)$ etc.

3.3.3 Verlustfunktion und Risiko

Wir haben oben in 3.3.2 gesehen, dass wir offenbar mehrere d's zur Wahl haben. Welches sollten wir nehmen? Dazu definiert man eine **Verlustfunktion** L (engl. Loss-Function). Z.B. beim Problem der *Schätzung* von μ bei der Normalverteilung:

$$L : \mathbb{R} \times \mathcal{A} \to \mathbb{R}. \qquad \text{(Estimation)}$$

Falls μ_0 das (uns (vorerst) unbekannte) richtige μ ist, dann ist $L(\mu_0, a)$ der Verlust, den wir machen, wenn wir Aktion a wählen. Man kann in Schätzungen z.B. eine quadratische Verlustfunktion nehmen:

$$L(\mu_0, a) = (\mu_0 - a)^2,$$

alternativ $L(\mu_0, a) = |\mu_0 - a|$. In einem *Testproblem* mit

$$L : \{\mathcal{H}_0, \mathcal{H}_1\} \times \{\text{sage}\,\mathcal{H}_0, \text{sage}\,\mathcal{H}_1\} \to \mathbb{R}. \qquad \text{(Test)}$$

kann man spezifizieren, dass ein Fehler 1. Art gravierender ist als ein Fehler 2. Art (oder umgekehrt): $L(\mathcal{H}_0, \text{sage}\,\mathcal{H}_0) = 0, L(\mathcal{H}_0, \text{sage}\,\mathcal{H}_1) = c_1, L(\mathcal{H}_1, \text{sage}\,\mathcal{H}_0) = c_2$,
$L(\mathcal{H}_1, \text{sage}\,\mathcal{H}_1) = 0$.

Wir werden jedoch nicht nur die gerade vorliegende Stichprobe bei der Wahl von $d \in \mathcal{D}$ einbeziehen, sondern einen erwarteten Verlust anschauen: Das Risiko $\mathcal{R}$ von Entscheidungsfunktion (oder Regel) d, wenn μ der richtige Erwartungswert ist, definieren wir als Abbildung

$$\mathcal{R} : \mathbb{R} \times \mathcal{D} \to \mathbb{R},$$

und zwar $\mathcal{R}(\mu, d) := E_\mu[L(\mu, d(X_1, \ldots, X_n))]$. Das war ein bisschen viel geschachtelte Definition. Wir betrachten als Beispiel die Schätzung von μ bei der $\mathcal{N}(\mu, \sigma^2)$-Verteilung.

Wir arbeiten mit einer quadratischen Verlustfunktion. Mögliche Entscheidungsfunktionen sind:

$$d_1(x_1, \ldots, x_n) = \overline{x},$$

oder ("I" für Ignorant, schaut nur gerade erste Zahl an)

$$d_I(x_1, \ldots, x_n) = x_1,$$

oder zum Beispiel auch ("SI" für Super-Ignorant, schaut gar keine Realisation an)

$$d_{SI}(x_1, \ldots, x_n) = 1.41421.$$

Wir berechnen jetzt die Risiken dieser Entscheidungsfunktionen.

3.3.4 Zulässigkeit

Wir haben anhand obiger drei Risiken gesehen, dass eine Entscheidungsfunktion durchs Band (im wahrsten Sinn des Wortes) schlechter sein kann als eine andere (d_I im Vergleich zu d_1). Deshalb definiert man: Eine Entscheidungsfunktion d ist **unzulässig**, wenn es eine andere Entscheidungsfunktion d^* gibt, sodass

$$\mathcal{R}(\mu, d^*) \leq \mathcal{R}(\mu, d) \; \forall \; \mu \in \mathbb{R}$$

$$\text{und } \exists \, \mu_*\text{:}$$

$$\mathcal{R}(\mu_*, d^*) < \mathcal{R}(\mu_*, d).$$

Wenn kein solches d^* existiert, dann nennen wir d **zulässig**. Welche der drei obigen Entscheidungsregeln ist/sind nicht zulässig?

3.4 Suffiziente Statistik

Wir wollen uns als StatistikerInnen bei Datenanalysen nicht unnötig einschränken lassen und definieren deshalb erstmal sehr frei:

Definition 3.1 [Statistik] *Sei $(x_1, x_2, \ldots, x_n)$ eine Stichprobe (gesamthaft aus dem $\mathbb{R}^n$). Wir definieren: Eine Statistik T ist eine beliebige (messbare) Funktion der Daten:*

$$\mathbb{R}^n \to V,$$

wobei V eine beliebige Teilmenge des $\mathbb{R}^k$ ist (indem sich ein zu schätzender Parameter befindet).

Ein Schätzer für den Mittelwert μ einer Normalverteilung ist also auch eine "Statistik". Wir haben in 3.3.3 gesehen, dass wir im Fall einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ sehr viele Möglichkeiten haben, z.B. μ zu schätzen (z.B. $\overline{x}, x_1, 1.41421$). Ein erstes Ziel ist jetzt, die Daten (gesamthaft im $\mathbb{R}^n$) so zu reduzieren, dass wir keine wesentliche Information für unser Problem (μ schätzen) verlieren. Mit dem Schätzer x_1 sagt uns bereits das Gefühl, dass wir da zu viel Information weggeschmissen haben. Wir definieren dazu:

Definition 3.2 [suffiziente Statistik] *Sei $(x_1, x_2, \ldots, x_n)$ eine Stichprobe aus einer beliebigen Verteilung P_θ mit unbekanntem Parameter θ. Eine Statistik T heisst suffizient, wenn*

$$P_\theta[(X_1, X_2, \ldots, X_n) \in A | T(X_1, X_2, \ldots, X_n) = t] = \lambda(t, A) \qquad \text{(suff)}$$

für alle $t, A \subseteq \mathbb{R}^n$; θ kommt auf der rechten Seite nicht mehr vor!

Diese Definition (v.a. die später folgende *minimal* suffiziente Statistik) erlaubt uns also, für ein statistisches Problem die Datenmenge gewaltig zu reduzieren, ohne Verlust an relevanter Information. Wir können die Datenmenge auch unverändert lassen: $T(x_1, \ldots, x_n) = (x_1, \ldots, x_n)$ ist suffizient. Warum?

Beispiel 1: Wir illustrieren diese Definition anhand der Bernoulli-Verteilung mit $\theta \in (0,1)$: $P_\theta[X_i = a] = \theta^a(1-\theta)^{1-a}$. Wir nehmen an, wir haben $(X_i)_{i=1}^n$ iid Be(θ) und $\mathbf{x} := (x_1, \ldots, x_n)$ ist eine Stichprobe hieraus. Wir wollen θ schätzen mit Hilfe von

$$T_1(x_1, \ldots, x_n) := \sum_{i=1}^n x_i,$$

oder mit einer Funktion hiervon (T_1/n). Ist dieses T_1 suffizient im Sinne von Definition 3.2? Wir überprüfen dies mit Hilfe eines speziellen $A := \{(x_1, \ldots, x_n)\}$ (allgemeinere A sind disjunkte Vereinigungen derartiger elementarer Bausteine und die bedingte Wahrscheinlichkeit $P[.|T(X_1, X_2, \ldots, X_n) = t]$ ist selber auch eine Wahrscheinlichkeit):

$$\begin{aligned}
P_\theta[(X_1, \ldots, X_n) \in A | T_1(X_1, \ldots, X_n) = t] = \\
&= \frac{P_\theta[\{(X_1, \ldots, X_n) \in A\} \cap \{T_1(X_1, \ldots, X_n) = t\}]}{P_\theta[T_1(X_1, \ldots, X_n) = t]} \\
&= \frac{P_\theta[\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n)\} \cap \{T_1(X_1, \ldots, X_n) = t\}]}{P_\theta[T_1(X_1, \ldots, X_n) = t]} \\
&= 1_{\{T_1(x_1,\ldots,x_n)=t\}}(\mathbf{x}) \frac{P_\theta[(X_1, \ldots, X_n) = (x_1, \ldots, x_n)]}{P_\theta[T_1(X_1, \ldots, X_n) = t]} \\
&= 1_{\{T_1(x_1,\ldots,x_n)=t\}}(\mathbf{x}) \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{P_\theta[T_1(X_1, \ldots, X_n) = t]} \\
&= 1_{\{T_1(x_1,\ldots,x_n)=t\}}(\mathbf{x}) \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\sum_{\{\mathbf{y}:T_1(\mathbf{y})=t\}} \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}} \\
&= 1_{\{T_1(x_1,\ldots,x_n)=t\}}(\mathbf{x}) \frac{\theta^t(1-\theta)^{n-t}}{\sum_{\{\mathbf{y}:T_1(\mathbf{y})=t\}} \theta^t(1-\theta)^{n-t}} \\
&= \frac{1_{\{T_1(x_1,\ldots,x_n)=t\}}(\mathbf{x})}{\sum_{\{\mathbf{y}:T_1(\mathbf{y})=t\}} 1} \\
&= \frac{1_{\{T_1(x_1,\ldots,x_n)=t\}}(\mathbf{x})}{\binom{n}{t}} \\
&=: \lambda(t, A).
\end{aligned}$$

Die Folge ist also: Wenn wir uns bei einer Stichprobe vom Umfang n aus einer Be(θ)-Verteilung nur für den Parameter θ interessieren (bisschen blödes Beispiel - was könnte uns denn sonst interessieren?), müssen wir nicht alle n Daten speichern, sondern nur die Summe T_1. Wir haben dabei keine relevante Information weggegeben.

Man kann sich leicht vorstellen, dass obige Rechnungen kompliziert werden können. Es gibt jedoch das nachfolgende Faktorisierungskriterium, mit Hilfe dessen wir durch Untersuchen der (gemeinsamen) Wahrscheinlichkeitsfunktion resp. Dichte bereits schliessen können, ob eine Statistik suffizient ist oder nicht. Wir erinnern uns von Kapitel 1.2.4, dass die gemeinsame Wahrscheinlichkeitsfunktion/Dichte im Fall von unabhängigen Zufallsgrössen (Stichproben!) wegen Formeln (I, III) einfach das Produkt der eindimensionalen Wahrscheinlichkeitsfunktionen bzw. Rand-Dichten ist.

Lemma 3.3 [Faktorisierungskriterium ("θ nur durch T an die Daten gebunden")] *Sei $(X_1, \ldots, X_n)$ eine iid Folge von Zufallsgrössen mit gemeinsamer Wahrscheinlichkeitsfunktion $p(x_1, \ldots, x_n; \theta)$ resp. Dichte $f(x_1, \ldots, x_n; \theta)$ (θ kann dabei auch aus dem $\mathbb{R} \times \mathbb{R}_+$ sein ((μ, σ^2) bei $\mathcal{N}(\mu, \sigma^2)$)). Dann ist T suffizient für θ genau dann wenn sich die Wahrscheinlichkeitsfunktion faktorisieren lässt in*

$$p(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n) g(T(x_1, \ldots, x_n), \theta)$$

resp. die Dichte faktorisieren lässt in

$$f(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n) g(T(x_1, \ldots, x_n), \theta).$$

Beweis Lemma 3.3 (diskreter Fall): Wir definieren $\mathbf{x} := (x_1, \ldots, x_n)$ und $\mathbf{X} := (X_1, \ldots, X_n)$.

"$\Rightarrow$": Sei T suffizient für θ, das heisst

$$P_\theta[\mathbf{X} \in A | T(\mathbf{X}) = t] = \lambda(t, A).$$

Dann gilt mit $T(\mathbf{x}) =: t$ und $A := \{\mathbf{x}\}$:

$$\begin{aligned} p(x_1, \ldots, x_n; \theta) := P_\theta[\mathbf{X} = \mathbf{x}] &= P_\theta[\{\mathbf{X} = \mathbf{x}\} \cap \{T(\mathbf{X}) = T(\mathbf{x})\}] \\ &= P_\theta[\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})] P_\theta[T(\mathbf{X}) = T(\mathbf{x})] \\ &= \lambda(t, \mathbf{x}) P_\theta[T(\mathbf{X}) = T(\mathbf{x})]. \end{aligned}$$

Wir haben damit p erfolgreich derart faktorisiert, dass θ nur noch an $T(\mathbf{x})$ gebunden vorkommt (in $P_\theta[T(\mathbf{X}) = T(\mathbf{x})]$).

"$\Leftarrow$": Sei jetzt im Gegenzug $p(x_1,\ldots,x_n;\theta) = h(x_1,\ldots,x_n)g(T(x_1,\ldots,x_n),\theta)$. Dann gilt (wieder mit speziellem A):

$$\begin{aligned}
P_\theta[\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t] &= \frac{1_{\{T(x_1,\ldots,x_n)=t\}}(\mathbf{x})P_\theta[(X_1,\ldots,X_n)=(x_1,\ldots,x_n)]}{P_\theta[T(X_1,\ldots,X_n)=t]} \\
&= \frac{1_{\{T(x_1,\ldots,x_n)=t\}}(\mathbf{x})p(x_1,\ldots,x_n;\theta)}{\sum_{\{\mathbf{x}':T(\mathbf{x}')=t\}} p(x_1',\ldots,x_n';\theta)} \\
&= \frac{1_{\{T(x_1,\ldots,x_n)=t\}}(\mathbf{x})h(x_1,\ldots,x_n)g(T(x_1,\ldots,x_n),\theta)}{\sum_{\{\mathbf{x}':T(\mathbf{x}')=t\}} h(x_1',\ldots,x_n')g(T(x_1',\ldots,x_n'),\theta)} \\
&= \frac{1_{\{T(x_1,\ldots,x_n)=t\}}(\mathbf{x})h(x_1,\ldots,x_n)g(t,\theta)}{\sum_{\{\mathbf{x}':T(\mathbf{x}')=t\}} h(x_1',\ldots,x_n')g(t,\theta)} \\
&= \frac{1_{\{T(x_1,\ldots,x_n)=t\}}(\mathbf{x})h(x_1,\ldots,x_n)}{\sum_{\{\mathbf{x}':T(\mathbf{x}')=t\}} h(x_1',\ldots,x_n')};
\end{aligned}$$

θ kommt nicht mehr vor.

□

Beispiel 2: Wir wollen das Faktorisierungskriterium anhand der Normalverteilung $\mathcal{N}(\mu,1)$ illustrieren. Die gemeinsame Dichte ist wegen Kapitel 1.2.4 (Formel III)

$$\begin{aligned}
f(x_1,\ldots,x_n;\mu,1) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\Big(\sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n x_i\mu + n\mu^2\Big)\right] \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\Big(\sum_{i=1}^n x_i^2\Big)\right] \exp\left[-\frac{1}{2}\Big(-2\sum_{i=1}^n x_i\mu + n\mu^2\Big)\right].
\end{aligned}$$

Offenbar ist hier wegen des Faktorisierungskriteriums bei *bekannter* Varianz (σ^2 muss nicht unbedingt gleich 1 sein, sondern kann auch jede beliebige andere, bekannte Zahl sein) die suffiziente Statistik für μ gleich $\sum_{i=1}^n x_i$. Wie sieht es aus, wenn wir beide Parameter (μ und σ^2) nicht kennen?

Beispiel 3: Normalverteilung mit zwei unbekannten Parametern $\mathcal{N}(\mu, \sigma^2)$: Die gemeinsame Dichte ist

$$\begin{aligned} f(x_1, \ldots, x_n; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\mu + n\mu^2\right)\right] \\ &=: g\big((\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i), (\mu, \sigma^2)\big). \end{aligned}$$

Jetzt sind wir nicht ganz so erfolgreich gewesen, aber immerhin konnten wir die Datenmenge auf zwei Statistiken reduzieren (oder eine zweidimensionale Statistik):

$$(\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i)$$

ist eine suffiziente Statistik für (μ, σ^2).

Wir haben festgehalten, dass beispielsweise die unveränderten Daten auch suffiziente Statistiken sind. Offenbar ist in Beispiel 2 die Summe der Stichprobenwerte suffizient und nur noch eine einzige Zahl. Man kann sich (vor allem wegen Beispiel 3) fragen, ob man die Daten nicht noch weiter reduzieren kann und immer noch eine suffiziente Statistik hat (die Antwort in Beispiel 3 ist übrigens klar NEIN!). Wir kommen damit zum **Konzept der minimal suffizienten Statistik**, welche wir in 4 Schritten einführen:

Schritt 1: Partition $\leftrightarrow$ Statistik

Zuerst ein bisschen Algebra: Wenn wir eine Stichprobe $\mathbf{x}$ (z.B. aus dem $\mathbb{R}^n$) und eine Statistik T (z.B. $T : \mathbb{R}^n \to \mathbb{R}$) haben, so ergibt dies eine Partition des **Stichprobenraums** K: wir fassen alle Elemente in einer Teilmenge zusammen, für die $T(\mathbf{x})$ denselben Wert (z.B. t) ergibt. Wir haben damit das Element K_t der Partition:

$$K_t := \{\mathbf{x} | T(\mathbf{x}) = t\}.$$

Wenn wir die Identität als Statistik nehmen ($id(\mathbf{x}) = \mathbf{x}$), so haben wir den Stichprobenraum auch in eine Partition zerlegt. Diese (feinste) Partition hat aber sehr viele Elemente (gleich viele wie der Stichprobenraum). Wir wollen jedoch auf sinnvolle Art und Weise eine Partition mit möglichst wenigen Elementen erhalten (möglichst grob).

Schritt 2: Konstruktion der Likelihood-Quotient-Partition

Dazu definieren wir ($p(\mathbf{x}, \theta)$ bezeichne entweder eine Wahrscheinlichkeitsfunktion oder eine Dichte an der Stelle $\mathbf{x}$ mit unbekannten Parametern θ)

$$D_0 := \{\mathbf{x} | p(\mathbf{x}, \theta) = 0 \, \forall \, \theta\};$$

es gilt dann $P_\theta[X \in D_0] = 0$ für alle θ. Dann definieren wir eine Relation $\sim$ über $K \backslash D_0$:

$$\mathbf{x} \sim \mathbf{y} \Leftrightarrow p(\mathbf{x}, \theta) = k(\mathbf{x}, \mathbf{y}) p(\mathbf{y}, \theta) \quad \text{mit} \quad k(\mathbf{x}, \mathbf{y}) \in (0, \infty), \tag{LQP}$$

oder anders formuliert:

$$\frac{p(\mathbf{x}, \theta)}{p(\mathbf{y}, \theta)}$$

hat kein θ mehr drin! Am Besten stellt man sich bei dieser ganzen Übung vor, man will den Mittelwert einer Normalverteilung schätzen und $\mathbf{x}$ und $\mathbf{y}$ sind Stichproben, welche das selbe arithmetische Mittel besitzen. Es gilt dann:

Lemma 3.4 $\sim$ *ist eine Äquivalenzrelation über* $K \backslash D_0$.

Beweis von Lemma 3.4:

□

Für ein $\mathbf{x}$ definiert man dann $D_{\mathbf{x}} := \{\mathbf{y} | p(\mathbf{y}, \theta) = k(\mathbf{y}, \mathbf{x}) p(\mathbf{x}, \theta) \; \forall \; \theta\}$. D_t, t in einer Indexmenge, seien die Äquivalenzklassen unter $\sim$. Wir nennen diese Partition die **Likelihood-Quotient-Partition LQP** (von $\sim$ erzeugt).

Schritt 3: LQP und Suffizienz

Wie kommen wir jetzt von der Partition zu einer (minimal) suffizienten Statistik?

In den jetzt folgenden Schritten setzen wir zuerst einfach voraus, dass die unten gewonnene (minimal) suffiziente Statistik auch messbar ist (und damit in der Tat überhaupt eine Statistik ist). Wir kommen am Schluss dieses Kapitels auf dieses Problem zurück.

Jedem D_t können wir ein Element $\mathbf{x}(t) \in D_t$ zuordnen (Auswahlaxiom!). Dann betrachten wir Gleichung (LQP). Für ein beliebiges Element $\mathbf{x}$ aus D_t gilt

$$p(\mathbf{x}, \theta) = k(\mathbf{x}, \mathbf{x}(t)) p(\mathbf{x}(t), \theta).$$

Wegen des Faktorisierungskriteriums ist damit $\mathbf{x}(t)$ suffizient für θ. Mehr, und das ist das Zentrale: **jede Statistik $T(\mathbf{x})$, welche die LQP erzeugt, ist suffiziente Statistik für θ.** Wir nennen diese Statistiken LQP-Statistiken.

Schritt 4: Minimalität

Wie steht es aber mit der Minimalität? Wir werden sagen, eine Statistik ist *minimal* suffizient, wenn die dazugehörige Partition (Schritt 1) die gröbste ist, welche wir unter den Partitionen finden, welche von suffizienten Statistiken hervorgebracht werden.

Satz 3.5 [über die minimale Suffizienz der LQP-Statistiken] *Die LQP-Statistiken sind minimal suffizient über $K \backslash D_0$.*

Beweis von Satz 3.5: Sei W eine beliebige, suffiziente Statistik. D_t^W sei die Partition, welche von W hervorgebracht wird. Wir zeigen, dass wir für ein $C^W \in D_t^W \backslash D_0$

immer ein C in der LQP finden, sodass $C^W \subseteq C$ (also LQP am gröbsten): Seien dazu $\mathbf{x}, \mathbf{y} \in C^W$. Wir haben damit (wegen der Suffizienz):

$$p(\mathbf{x}, \theta) = h(\mathbf{x})g(W(\mathbf{x}), \theta)$$

und

$$p(\mathbf{y}, \theta) = h(\mathbf{y})g(W(\mathbf{y}), \theta).$$

Damit gilt aber weil $W(\mathbf{x}) = W(\mathbf{y})$, dass

$$p(\mathbf{x}, \theta) = \frac{h(\mathbf{x})}{h(\mathbf{y})} p(\mathbf{y}, \theta).$$

Damit muss aber $x \sim y$ gelten und damit sind x und y auch in einem gemeinsamen C aus der LQP.

□

Zusammenfassend haben wir zur Konstruktion von (minimal) suffizienten Statistiken folgendes einfaches Rezept (vgl. Aufgabe in Übungen):

$$\frac{p(\mathbf{x}, \theta)}{p(\mathbf{y}, \theta)}$$

muss einfach θ-frei sein. Dies sieht seltsam aus und sollte gleich mit einem (bereits bekannten) Beispiel illustriert werden.

Zur *Messbarkeit* obiger Statistik: Wir haben oben bei der Konstruktion und Beweisführung einfach vorausgesetzt, die gewonnene Funktion der Daten sei automatisch eine messbare Funktion und damit eine Statistik. Dies kann man aber in dieser Allgemeinheit genau genommen nicht einfach voraussetzen. Hingegen können wir anhand der praktischen Beispiele jeweils immer überprüfen, ob die so gewonnene Funktion messbar ist (zum Beispiel $\bar{X}$) und damit auch wirklich eine Statistik ist. Dank der Existenz einer mit (LQP) gewonnenen Statistik im konkreten Beispiel haben wir dann für das jeweils vorliegende Beispiel auch die minimal suffiziente Statistik, weil alle minimal suffizienten Statistiken die gleiche Partition erzeugen.

Beispiel 3 revisited: Normalverteilung mit zwei unbekannten Parametern $\mathcal{N}(\mu, \sigma^2)$: Der Quotient der Dichten ist

$$\begin{aligned}\frac{f(x_1,\ldots,x_n;\mu,\sigma^2)}{f(y_1,\ldots,y_n;\mu,\sigma^2)} &= \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}} \\ &= \exp\big[-\frac{1}{2\sigma^2}(\sum_{i=1}^n (x_i-\mu)^2 - \sum_{i=1}^n (y_i-\mu)^2)\big] \\ &= \exp\big[-\frac{1}{2\sigma^2}(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2 + 2n\mu(\overline{y}-\overline{x}))\big]\end{aligned}$$

Aus einfachen, algebraischen Überlegungen folgt jetzt, dass dieser Ausdruck nur dann θ-frei ist (hier (μ, σ^2)-frei), wenn

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$$

und

$$\overline{y} = \overline{x}.$$

So wird hier also der Stichprobenraum partitioniert. Die Statistiken

$$(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)$$

sind somit minimal suffizient.