

Statistische Methoden

Dr. C.J. Luchsinger

5 Schätztheorie und Konfidenzintervalle

Literatur Kapitel 5

- * Lindgren: Kapitel 8
- * Cartoon Guide: Kapitel 7
- * Krengel: § 4, 13
- * Stahel: Kapitel 7, 9

Wir wollen die beiden Gebiete "Testen" und "Schätzen" kurz von der Wahrscheinlichkeitstheorie her anschauen. Es geht dabei um die Frage, welche beiden zentralen Hilfsmittel aus der Wahrscheinlichkeitstheorie wo eingesetzt werden. Die beiden Theoreme sind LLN (Satz 1.24) und CLT (Theorem 1.25):

- * In Kapitel 4 haben wir zuerst exakte Methoden entwickelt. Danach folgten die χ^2 -Tests und die approximativen Verfahren aus 4.5.4. Betrachten wir exemplarisch die erste Teststatistik aus 4.5.4:

$$\frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} = \frac{\sum x_i - n\mu_0}{\sigma \sqrt{n}}.$$

Wir wollen am Schluss für die Anwendungen einen Ausdruck haben, welcher insofern stabil ist, als dass die Varianz weder kollabiert (gegen 0 geht) noch über jede Schranke wächst. Mit der Normierung mit \sqrt{n} im Nenner im CLT wird genau dies erreicht. Zudem erhalten wir unter sehr allgemeinen Bedingungen immer eine Normalverteilung als Limesverteilung. Diese ist tabelliert (in Büchern oder Rechenlibraries) und wir können dort mindestens asymptotisch kritische Werte ablesen (1.64, 1.96). In der Praxis ist der CLT zentral wichtig für approximative Test-Verfahren, weil man die genaue Verteilung oft nicht theoretisch berechnen kann.

- * In der Schätztheorie von Kapitel 5 wollen wir jedoch möglichst einen Ausdruck, welcher zu einem Punkt kollabiert. Die Varianz soll gegen 0 gehen. Deshalb ist der LLN hier zentral wichtiges Arbeitsinstrument.

5.1 Schätztheorie

Was wir in 5.1 erarbeiten werden, ist vielen StudentInnen vielleicht schon einmal durch den Kopf gegangen - es geht jetzt vor allem darum, diese Gedanken zu ordnen. Der Aufbau ist wie folgt:

5.1.1 Motivierendes Beispiel

5.1.2 Definition Schätzer

5.1.3 Anforderungen an Schätzer

5.1.4 Allgemeine Schätzverfahren: Maximum Likelihood und Momentenmethode

5.1.5 Die Schätzung von σ^2 und σ im Modell $\mathcal{N}(\mu, \sigma^2)$

5.1.6 Die Cramer-Rao-Schranke

5.1.7 Abschliessende Bemerkungen zum Schätzproblem

5.1.1 Motivierendes Beispiel; siehe Vlsg WTS

5.1.2 Definition Schätzer; siehe Vlsg WTS

Definition 5.1 [Schätzer, engl. Estimator] *Ein Schätzer $\hat{\mu}_n := \hat{\mu}_n(x_1, \dots, x_n)$ für einen unbekannt Parameter μ ist eine beliebige (messbare) Funktion der Daten. Insbesondere kann die Schätzfunktion die Daten vollständig ignorieren.*

Es sei noch darauf hingewiesen, dass, Bezug nehmend auf 3.3 "Mathematische Formalisierung", ein Schätzer eine Entscheidungsfunktion ist.

5.1.3 Anforderungen an Schätzer; siehe Vlsg WTS

Definition 5.2 [erwartungstreu (unverfälscht), engl. unbiased; Bias] *Ein Schätzer $\hat{\mu}_n$ für μ heisst erwartungstreu oder unverfälscht, wenn*

$$E_{\mu}[\hat{\mu}_n] = \mu$$

für alle μ im Parameterraum. Ansonsten spricht man davon, dass der Schätzer einen Bias b hat; wir definieren

$$b := E_{\mu}[\hat{\mu}_n - \mu].$$

Bemerkung zu Definition 5.2: Wie in Kapitel 4.3 bereits angekündigt, hat dieses "bias" *nichts* mit dem "unbiased" bei den Tests (UMPU) zu tun!

Definition 5.3 [Konsistenz] *Ein Schätzer $\hat{\mu}_n$ für μ ist konsistent, wenn für alle μ im Parameterraum, $\epsilon > 0$ gilt:*

$$\lim_{n \rightarrow \infty} P_\mu [|\hat{\mu}_n - \mu| > \epsilon] = 0.$$

Definition 5.4 [kleinste Varianz] *Ein Schätzer $\hat{\mu}_n$ für μ hat kleinste Varianz für n fix, wenn die Varianz von $\hat{\mu}_n$ unter P_μ minimal ist für alle μ im Parameterraum.*

Wir werden in 5.1.6 (Cramer-Rao-Schranke) sehen: Wenn wir eine Stichprobe aus einer Normalverteilung haben, dann hat unter den erwartungstreuen Schätzern das arithmetische Mittel die kleinstmögliche Varianz.

Definition 5.5 [minimaler Mean-Square-Error (MSE)] *Ein Schätzer $\hat{\mu}_n$ für μ minimiert den Mean-Square-Error (MSE) für n fix, wenn*

$$MSE(\hat{\mu}_n, \mu) := E_\mu[(\hat{\mu}_n - \mu)^2]$$

minimal ist für alle μ im Parameterraum.

Auf die Schnelle könnte man meinen, MSE sei die Varianz des Schätzers. Dies stimmt jedoch nur, wenn der Schätzer erwartungstreu ist (Bias $b = 0$), wie wir auch aus folgendem Lemma schliessen können:

Lemma 5.6 *[$MSE = V + b^2$] Mit obigen Bezeichnungen gilt:*

$$MSE(\hat{\mu}_n, \mu) = V[\hat{\mu}_n] + b^2.$$

Daneben gibt es noch die schwieriger zu definierende Eigenschaft der **Robustheit**. Es geht dabei um die Anforderung, dass falsche Messungen die Schätzungen für μ nicht stark beeinflussen dürfen. Wir sind dann aber eigentlich gar nicht mehr z.B. im Modell $\mathcal{N}(\mu, \sigma^2)$. Der Median ist ein sehr robuster Schätzer. Betrachten wir dazu einen Vergleich von arithmetischem Mittel und Median. Wenn man in (x_1, \dots, x_n) selbst bei grossem n auch nur einen Wert total verfälscht ("nach 10^9 und mehr schickt"), dann wird das arithmetische Mittel beliebig falsch. Hingegen ändert sich der Median kaum. In der Tat kann man beim Median bis zu $< 50\%$ aller Daten total verfälschen; die Schätzung hat immer noch "etwas mit dem richtigen Wert zu tun". Dieses "etwas mit dem richtigen Wert zu tun" haben lässt sich mathematisch sauber ausformulieren (Bruchpunkt), wir verzichten hier aus Zeitgründen darauf. Das Forschungsgebiet heisst *robuste Statistik*, ist sehr wichtig, komplex, kompliziert, aktuell und benötigt sehr gute Vorkenntnisse in *reiner* Mathematik.

Abschliessend kann man sagen, dass bei $\mathcal{N}(\mu, 1)$ das arithmetische Mittel bei 100%-Datenqualität (keine Verfälschungen) der beste Schätzer ist (unverfälscht, konsistent, kleinste Varianz unter den unverfälschten Schätzern, minimaler MSE). σ muss dabei nicht 1 sein - σ muss nicht mal bekannt sein.

5.1.4 Allgemeine Schätzverfahren: Maximum Likelihood und Momentenmethode; siehe Vlsg WTS

Maximum Likelihood Estimator MLE

In Kapitel 4 haben wir bei der Behandlung des Lemmas von Neyman-Pearson (NP) bereits kurz den Begriff der "Likelihood" erwähnt (den Test nach NP nennt man auch Likelihood-Ratio-Test). Dort haben wir den Quotienten der Dichten (oder der Wahrscheinlichkeitsfunktionen) angeschaut. Die Daten waren dabei fest, der Parameter war variabel, einmal z.B. μ_0 , dann μ_1 . Dies ist genau die Likelihood:

Definition 5.7 [Likelihood] *Wenn wir eine Dichtefunktion $f_\mu(x_1, \dots, x_n)$ oder Wahrscheinlichkeitsfunktion $p_\theta(x_1, \dots, x_n)$ als Funktion des Parameters μ resp. θ auffassen bei konstanten Daten, dann spricht man von der Likelihood.*

Man kann auch sagen, dass der Begriff der Likelihood aus der Not entstanden ist: wir haben ja nur die Daten (die können wir nicht mehr variieren) und gehen mal davon aus, dass die Daten z.B. aus einer Normalverteilung stammen. Dann können wir nur noch die Parameter variieren und schauen, was passiert.

Wir haben noch nicht die *Methode*:

$$\hat{\mu}_n^{MLE} := \operatorname{argmax}_{\mu \in \mathbb{R}} f_{\mu}(x_1, \dots, x_n),$$

analog im Fall von Wahrscheinlichkeitsfunktionen.

Wie ist im Fall $\mathcal{N}(\mu, 1)$ der MLE?

$$f_{\mu}(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp \left[-\frac{1}{2} \left(\sum_{i=1}^n x_i^2 - 2\mu n\bar{x} + n\mu^2 \right) \right]$$

Da wir nur das argmax suchen, können wir gerade so gut den Ausdruck

$$\exp \left[-\frac{1}{2} \left(\sum_{i=1}^n x_i^2 - 2\mu n\bar{x} + n\mu^2 \right) \right]$$

maximieren. Da der Logarithmus die Ordnung nicht verändert (und wir nur argmax suchen), können wir gerade so gut den Ausdruck

$$-\frac{1}{2} \left(\sum_{i=1}^n x_i^2 - 2\mu n\bar{x} + n\mu^2 \right)$$

maximieren. Dieser Schritt (Logarithmieren) ist zentral wichtig in vielen Rechnungen der Statistik, um das Leben einfacher zu machen. Wir müssen letztendlich den Ausdruck

$$-2\mu\bar{x} + \mu^2$$

minimieren, Daten fest, μ variabel:

Dieses Resultat gilt übrigens für alle σ .

In einer Aufgabe müssen Sie den MLE bei der Exponentialverteilung finden. Analog verfährt man mit diskreten Verteilungen.

Wir wissen noch nicht, weshalb diese Methode gut sein soll. Eine (teilweise) Antwort liefert folgender

Satz 5.8 [Rechtfertigung für den MLE; interpretationsbedürftig] Sei $p_\theta(x), \theta$ im Parameterraum, eine Familie von Wahrscheinlichkeitsfunktionen (oder Dichten). Der richtige, unbekannte Parameter sei θ_0 . Für alle θ im Parameterraum gelte $E_{\theta_0}[|\log p_\theta(X)|] < \infty$. Dann gilt:

$$J := E_{\theta_0}[\log p_{\theta_0}(X)] \geq E_{\theta_0}[\log p_\theta(X)] =: I \quad (J \geq I)$$

für alle θ im Parameterraum.

Beweis von Satz 5.8: Als kleine Hausaufgabe (HA) zeige man, dass für alle $x > 0$ gilt:

$$x \log x \geq 2(x - \sqrt{x}). \quad (\text{HA})$$

Wir vereinbaren folgende Notation:

$$h(x) := p_{\theta_0}(x), \quad q(x) := p_\theta(x).$$

Damit ist also wegen Lemma 1.17 a) zu zeigen, dass (Notation mit Integralen für den stetigen Fall, diskret mit Summen als analoge Hausaufgabe)

$$\int \log(h(x))h(x)dx \geq \int \log(q(x))h(x)dx.$$

Wir haben (man beachte (HA) sowie h und q sind Dichten, integrieren also auf 1)

$$\begin{aligned} J - I &:= \int \log(h(x))h(x)dx - \int \log(q(x))h(x)dx \\ &= \int \log\left(\frac{h(x)}{q(x)}\right)h(x)dx = \int \log\left(\frac{h(x)}{q(x)}\right)q(x)\frac{h(x)}{q(x)}dx \\ &\geq \int q(x)2\left(\frac{h(x)}{q(x)} - \sqrt{\frac{h(x)}{q(x)}}\right)dx \\ &= 2\left(1 - \int \sqrt{h(x)q(x)}dx\right) = 1 - 2 \int \sqrt{h(x)q(x)}dx + 1 \\ &= \int [q(x) - 2\sqrt{h(x)q(x)} + h(x)]dx \\ &= \int [\sqrt{q(x)} - \sqrt{h(x)}]^2 dx \geq 0. \end{aligned}$$

□

Interpretation von Satz 5.8: Formel ($J \geq I$) sagt aus, dass im Ausdruck

$$E_{\theta_0}[\log p_{\theta}(X)]$$

bei variablem θ in θ_0 ein Maximum erreicht wird. Wir suchen dieses θ_0 und man könnte jetzt versucht sein, einfach diesen Ausdruck nach θ abzuleiten, was wegen Satz 5.8 einfach θ_0 (und keine konkrete Zahl) herausspucken würde. Aber: die ML-Methode arbeitet mit dem Ausdruck

$$\log p_{\theta}(x_1, \dots, x_n)$$

und maximiert diesen bezüglich θ mit Daten x_1, \dots, x_n . Was geschieht hierbei, wenn wir auf die Ebene der Zufallsgrößen wechseln? Wir können obigen Ausdruck gleich noch durch n teilen, das "argmax" bleibt gleich; wir erhalten:

$$\frac{1}{n} \log p_{\theta}(X_1, \dots, X_n) = \frac{1}{n} \log \prod_{i=1}^n p_{\theta}(X_i) = \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i).$$

Wegen LLN konvergiert dieser Ausdruck aber gegen $E_{\theta_0}[\log p_{\theta}(X)]$ (im Sinne von Definition 1.23), wenn θ_0 der richtige Parameter der Zufallsgrößen X_i ist. Wir maximieren also auf der Ebene der $\log p_{\theta}(x_1, \dots, x_n)$ statt der $E_{\theta_0}[\log p_{\theta}(X)]$, wissen aber, dass wegen des LLN die ersten Ausdrücke wenigstens gegen die theoretischen Ausdrücke konvergieren. Dazu eine kleine Skizze:

Unter nicht allzu starken Voraussetzungen an die Wahrscheinlichkeitsfunktionen/Dichten und den Parameterraum lässt sich übrigens auch zeigen, dass im Sinne von Definition 1.23 folgende Konvergenz vorhanden ist:

$$\hat{\theta}_n^{MLE} \longrightarrow \theta_0, \quad n \rightarrow \infty.$$

Momentenmethode

Neben der ML-Schätzmethode gibt es unter anderem auch die Momentenmethode (eine weitere, Ordinary Least Squares (OLS, Kleinste Quadrat Schätzung), werden wir in Kapitel 7 "Regression" noch kennenlernen). Dies sind *allgemeine* Schätzverfahren - allgemein in dem Sinne, dass Sie für die verschiedensten Modelle ein allgemein einsetzbares Verfahren anbieten. Das Prinzip der Momentenmethode ist sehr einfach und wir sind alle sicher selber schon auf diese Idee gekommen:

Wenn eine Verteilungsfamilie k zu bestimmende, unbekannte Parameter hat (z.B. $\mathcal{N}(\mu, \sigma^2)$ mit $k = 2$), so wird man einfach k Momente (das sind die $E[X^j]$, j -tes Moment) der theoretischen und empirischen Verteilung gleichsetzen und die sich ergebenden Gleichungssysteme lösen:

$$E[X^j] = \frac{1}{n} \sum_{i=1}^n x_i^j, \quad 1 \leq j \leq k.$$

Gleichwertig kann man wegen der Bedeutung von $E[(X - \mu)^2]$ statt $E[X^2] = \frac{1}{n} \sum_{i=1}^n x_i^2$ gleich $E[(X - \mu)^2] = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ setzen.

Aus statistischen Gründen nimmt man die ersten Momente und nicht etwa z.B. bei der Normalverteilung das 10-te und 11-te Moment, da sonst allfällige Verunreinigungen der Daten zu stark in's Gewicht fallen würden (fehlende Robustheit!).

Berechnen Sie mit der Momentenmethode in der Stunde mit Hilfe von $E[X]$ und $E[X^2]$ Schätzungen von μ und σ^2 im Fall der Normalverteilung.

In einer Aufgabe ist ein (nicht mehr triviales) Beispiel mit der Gamma-Verteilung zu rechnen.

Wir haben am Anfang von Kapitel 5 darauf aufmerksam gemacht, dass der LLN in diesem Kapitel zentral wichtig ist (im Kapitel 4 eher der CLT). Mit dem LLN folgt: Die Momentenmethode liefert konsistente Schätzer, sobald die jeweiligen Momente existieren. Wer den LLN anschaut, ist vielleicht irritiert, weil dort ja nur die X^1 betrachtet werden. Aber falls z.B. $E[X^2] < \infty$, können wir mit $Y_i := X_i^2$ und $y_i := x_i^2, 1 \leq i \leq n$, den LLN wieder auf die Folge der Y_i anwenden (analog bei höheren Momenten).

* *

Wir wollen hier noch die Gelegenheit nutzen, um beim Schätzer "mit der Lupe" doch noch genauer hinzuschauen. Wegen des LLN (oder anderer Beweistechniken) konvergieren gutgewählte Schätzer gegen einen Punkt (konsistent). Aber wie sieht es genau aus? Wir können wegen des CLT z.B. im Fall des ersten Momentes und dessen Schätzung angeben (2. Moment muss existieren):

$$\sqrt{n}(\bar{X} - \mu) \rightarrow \mathcal{N}(0, \sigma^2)$$

im Sinne von Theorem 1.25. Wenn diese Eigenschaft

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow \mathcal{N}(0, \tau^2)$$

mit einer gewissen Varianz τ^2 gegeben ist, sagt man, der Schätzer $\hat{\theta}$ für θ ist asymptotisch normalverteilt. Achtung: Die Varianz des Schätzers geht gegen 0 und wir müssen mit \sqrt{n} "aufblähen", um diese Normalverteilung zu erhalten! Wenn $E[X^{2k}]$ existiert, so konvergieren alle empirischen Momente bis zur Potenz k derart gegen die theoretischen Momente, dass sie asymptotisch normalverteilt sind. Bei den MLE gibt es gewisse Regularitätsbedingungen an Wahrscheinlichkeitsfunktionen/Dichten und an den Parameterraum, welche garantieren, dass MLE auch asymptotisch normalverteilt ist.

5.1.5 Die Schätzung von σ^2 und σ im Modell $\mathcal{N}(\mu, \sigma^2)$; siehe Vlsg WTS

5.1.6 Die Cramer-Rao-Schranke

Wenn wir einen erwartungstreuen Schätzer $\hat{\theta}$ für einen Parameter θ haben, so ist es sehr gut, wenn auch noch die Varianz des Schätzers klein ist. Wegen der Ungleichung von Bienayme-Tschebyschew (vgl. 1.5.1) ist dann auch der Ausdruck

$$P[|\hat{\theta} - \theta| > \epsilon] \quad \left(\leq \frac{1}{\epsilon^2} V[\hat{\theta}] \right)$$

klein ($\epsilon > 0$ vorgegeben). Im Modell $\mathcal{N}(\mu, \sigma^2)$ haben wir mit \bar{x} einen nach fast allen denkbaren Kriterien (Ausnahme Robustheit) sehr guten Schätzer. Man kann sich aber fragen, ob nicht sehr intelligente Menschen einen noch besseren Schätzer basteln könnten. Suffizienzüberlegungen sagen uns, dass das arithmetische Mittel alle Informationen für dieses Problem noch besitzt. Ein Einbezug z.B. von $x_{(1)}$ oder Artverwandte ist nicht sinnvoll.

Vielleicht gibt es aber eine sehr komplizierte (messbare) Funktion

Superfunction(\bar{x})

(von ganz \bar{x}), welche noch besser ist? Die Antwort ist, zumindest was die Varianz des Schätzers anbelangt, klipp und klar: **NEIN!** (and we're gonna prove it...)

Satz 5.9 [Cramer-Rao-Schranke] Sei $\hat{\theta}$ ein erwartungstreuer Schätzer für θ . Dann gilt:

$$V_{\theta}[\hat{\theta}] \geq \frac{1}{I_{\theta}}, \quad (\text{CR} \neq \text{ung})$$

wobei (Formulierung mit stetiger Zufallsgrösse; diskret analog)

$$I_{\theta} := \int \left(\frac{\partial}{\partial \theta} \log(f(\mathbf{x}, \theta)) \right)^2 f(\mathbf{x}, \theta) d\mathbf{x}.$$

Wir fordern dazu (Regularity): 1. Der Wertebereich der Zufallsgrösse X darf nicht von θ abhängig sein und 2. folgende zwei Regularitätsbedingungen müssen erfüllt sein: für gegebenes θ muss eine kleine Nachbarschaft N_{θ} existieren, sodass

$$\int \sup_{\psi \in N_{\theta}} \left| \frac{\partial}{\partial \psi} f(\mathbf{x}, \psi) \right| d\mathbf{x} < \infty \quad \text{und} \quad \int \sup_{\psi \in N_{\theta}} \left| \hat{\theta}(\mathbf{x}) \frac{\partial}{\partial \psi} f(\mathbf{x}, \psi) \right| d\mathbf{x} < \infty.$$

Beweis von Satz 5.9: Wir definieren die Zufallsgrösse

$$S(\theta) := \frac{\partial}{\partial \theta} \log(f(\mathbf{X}, \theta))$$

und berechnen unter Verwendung von (Regularity) deren Erwartungswert:

$$E_\theta[S(\theta)] = \int \left[\frac{\partial}{\partial \theta} \log(f(\mathbf{x}, \theta)) \right] f(\mathbf{x}, \theta) d\mathbf{x} = \int \frac{\partial}{\partial \theta} f(\mathbf{x}, \theta) d\mathbf{x} = \frac{\partial}{\partial \theta} \int f(\mathbf{x}, \theta) d\mathbf{x} = \frac{\partial}{\partial \theta} 1 = 0.$$

Für einen (vorerst nicht unbedingt erwartungstreuen) beliebigen Schätzer $\hat{\theta}$ gilt (unter Verwendung von Regularity):

$$\begin{aligned} \text{Cov}_\theta(S(\theta), \hat{\theta}) &= E_\theta[S(\theta)\hat{\theta}] - E_\theta[S(\theta)]E_\theta[\hat{\theta}] = E_\theta[S(\theta)\hat{\theta}] \\ &= \int \hat{\theta}(\mathbf{x}) \frac{\partial}{\partial \theta} \log(f(\mathbf{x}, \theta)) f(\mathbf{x}, \theta) d\mathbf{x} \\ &= \int \hat{\theta}(\mathbf{x}) \frac{\partial}{\partial \theta} f(\mathbf{x}, \theta) d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \int \hat{\theta}(\mathbf{x}) f(\mathbf{x}, \theta) d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} E_\theta[\hat{\theta}(\mathbf{X})] = \frac{\partial}{\partial \theta} (\theta + E_\theta[\hat{\theta}(\mathbf{X})] - \theta) \\ &= \frac{\partial}{\partial \theta} (\theta + b) = 1 + b'. \end{aligned}$$

Dabei bezeichnen wir mit b, b' den Bias von $\hat{\theta}$ respektive seine Ableitung nach θ . Wegen Lemma 1.21 a) haben wir

$$\sqrt{V_\theta[S(\theta)]V_\theta[\hat{\theta}]} \geq \text{Cov}_\theta(S(\theta), \hat{\theta}) = 1 + b'.$$

Damit erhalten wir die eigentliche, allgemeine Ungleichung von Cramer-Rao:

$$V_\theta[\hat{\theta}] \geq \frac{(1 + b')^2}{V_\theta[S(\theta)]}$$

Mit $b = 0$ erhalten wir wegen

$$V_\theta[S(\theta)] = E_\theta[S(\theta)^2] = \int \left(\frac{\partial}{\partial \theta} \log(f(\mathbf{x}, \theta)) \right)^2 f(\mathbf{x}, \theta) d\mathbf{x} \quad (= I_\theta)$$

die Schranke aus Satz 5.9.

□

In der Vorlesung: Schranke für μ in $\mathcal{N}(\mu, \sigma^2)$.

In der Vorlesung: Schranke für $\theta := 1/\lambda$ in $\text{Exp}(\lambda)$.

5.1.7 Abschliessende Bemerkungen zum Schätzproblem

Es stellt sich immer die Frage, welchen Schätzer man einsetzen soll. Wichtig ist sicher die Forderung nach Konsistenz. Früher wurde immer die Unverfälschtheit gefordert, dieses Kriterium ist nicht mehr so en vogue. Sinnvoll ist sicher die Minimierung des MSE. Heutzutage wird auch mehr auf Robustheit geachtet.

Bei der Frage, welchen Schätzer man einsetzen soll, muss man sich folgende Fragen stellen:

“Wozu wird der Schätzer gebraucht?”

“Was soll mit dem Schätzer gemessen werden?”

Welches Kriterium auch immer gewählt wird: wichtig ist, dass man angibt, nach welcher Methode ein Parameter geschätzt wird (und welche Annahmen bei der Modellbildung gemacht werden). Dies ist sowieso bei allen statistischen Methoden zentral.

Hat man Abhängigkeitsstrukturen in den Daten (Zeitreihen), so müssen obige Schätzvorschläge für Lage- und Varianzparameter oft durch andere Schätzer ersetzt werden. Auch in Kapitel 7 (Regression) werden wir (zum Teil) andere Schätzer vorziehen.

Andere Vorschläge für Schätzer müssen immer zuerst auf obige Kriterien hin untersucht werden.

5.2 Konfidenzintervalle (Intervall-Schätzer)

5.2.1 Was ist ein Konfidenzintervall (KI) - was ist es *nicht*

Was ist es *nicht*: In den Nachrichten und wissenschaftlichen Publikationen liest man oft Sätze der Art (Zahlen frei erfunden): "Aufgrund einer Befragung mit einer Stichprobe von 10'000 Personen kam man zum Schluss, dass der Anteil der Anhänger von Bundeskanzler Müller mit 95 % Wahrscheinlichkeit in einem Konfidenzintervall von [46%, 48%] liegt."

Was ist hier falsch?

Definition 5.10 [Konfidenzintervalle] *Ein Konfidenzintervall KI für θ mit Konfidenzkoeffizient $(1-\alpha)$ ist eine zufällige Teilmenge des Parameterraums mit der Eigenschaft, dass*

$$P_{\theta}[\theta \in KI] = 1 - \alpha$$

für alle θ des Parameterraums (z.B. $\forall \theta \in \mathbb{R}$).

Bemerkungen zu Definition 5.10: Im Teil "was ist es *nicht*" haben wir bereits betont, dass θ nicht zufällig ist. Das Konfidenzintervall KI ist zufällig (*vor* der Realisation). Wenn danach die Realisation mit konkreten Zahlen vorliegt, sollte man Formulierungen brauchen wie: "[46%, 48%] ist eine Realisation eines 95 % Konfidenzintervalles". Aber auch vor der Realisation sollte man nicht sagen: " θ liegt mit Wahrscheinlichkeit $1 - \alpha$ in KI", sondern "KI deckt mit Wahrscheinlichkeit $1 - \alpha$ den Parameter θ ab" - um zu betonen, dass der Zufall (die Bewegung) in KI liegt und nicht in θ .

5.2.2 "Umgestülpte" zweiseitige Tests als optimale KI

Manche StudentInnen haben bei der Besprechung der zweiseitigen Tests vielleicht gespürt, dass diese etwas mit KI zu tun haben. KI sind umgangssprachlich formuliert so etwas wie "umgestülpte" zweiseitige Tests. Auch das " $1 - \alpha$ " (bei KI) im Gegensatz zu " α " (bei Tests) unterstreicht diese Tatsache. Dieses "Umstülpen" hat mathematisch einen edler klingenden Namen: *Dualitätsprinzip* (Test \leftrightarrow KI).

Frage an's Publikum: bei den Tests hatten wir die vollrandomisierten Tests. Gibt es auch so etwas bei den KI?

Wir zeigen jetzt die Konstruktion von KI aus zweiseitigen Tests:

Folgende Skizze illustriert die Dualität, welche wir auf der folgenden Seite formalisieren werden:

Wir haben folgende zweiseitige Testsituation: $\mathcal{H}_0 : \theta = \theta_0$, $\mathcal{H}_1 : \theta \neq \theta_0$. Sei $d(\mathbf{x})$ ein statistischer Test der Grösse α (stetiger Fall; $d(\mathbf{x}) = 0$ heisst \mathcal{H}_0 -Hypothese annehmen und $d(\mathbf{x}) = 1$ heisst \mathcal{H}_1 -Hypothese annehmen). Wir definieren mit

$$A_{\theta_0} := \{\mathbf{x} | d(\mathbf{x}) = 0\}$$

den Bereich, wo wir die \mathcal{H}_0 -Hypothese mit Daten \mathbf{x} annehmen. Jetzt definieren wir mit

$$KI_{\alpha}(\mathbf{x}) := \{\theta | \mathbf{x} \in A_{\theta}\}$$

die Menge aller θ , die wir als Nullhypothese bei Beobachtungen \mathbf{x} angenommen hätten.

* **Behauptung I:** dieses KI_{α} hat Konfidenzkoeffizient $1 - \alpha$:

$$P_{\theta}[\theta \in KI_{\alpha}(\mathbf{X})] = P_{\theta}[\mathbf{X} \in A_{\theta}] = 1 - \alpha.$$

□

* **Behauptung II:** dieses KI_{α} ist, wenn aus einem UMPU-Test konstruiert, in einem gewissen Sinne optimal (das Pendant zu Uniformly Most Powerful Unbiased), genauer: Sei KI'_{α} ein weiteres KI mit Konfidenzkoeffizient $1 - \alpha$, welches *nicht* aus einem UMPU-Test, sondern aus einem anderen unverfälschten Test konstruiert wurde. Wenn θ_0 richtig ist und $\theta \neq \theta_0$, dann gilt

$$P_{\theta_0}[\theta \in KI_{\alpha}] \leq P_{\theta_0}[\theta \in KI'_{\alpha}].$$

Dies ist wirklich gut: wenn θ ja nicht richtig ist, dann wollen wir es möglichst nicht im KI haben. Beweis:

$$\begin{aligned} P_{\theta_0}[\theta \in KI_{\alpha}] &= P_{\theta_0}[\mathbf{X} \in A_{\theta}] = 1 - P_{\theta_0}[\mathbf{X} \notin A_{\theta}] \leq 1 - P_{\theta_0}[\mathbf{X} \notin A'_{\theta}] \\ &= P_{\theta_0}[\mathbf{X} \in A'_{\theta}] = P_{\theta_0}[\theta \in KI'_{\alpha}]. \end{aligned}$$

Die Ungleichung $P_{\theta_0}[\mathbf{X} \notin A_{\theta}] \geq P_{\theta_0}[\mathbf{X} \notin A'_{\theta}]$ gilt, weil wir A_{θ} aus einem UMPU gewonnen haben.

□

Als kleine, freiwillige Hausaufgabe überlege man sich, dass man zu jedem KI eine analoge Testsituation finden kann. Zudem ist die Forderung nach unbiased sinnvoll: wenn sie nicht gilt hat man ein KI, bei dem "falsche" θ eher im KI sind als das richtige θ_0 . Zusammengefasst: KI_{α} (aus UMPU) ist das beste KI unter keinen *relevanten* Einschränkungen.

5.2.3 Wichtige Beispiele für KI

Wir werden jetzt die theoretischen Überlegungen aus 5.2.2 auf konkrete Situationen anwenden. Man beachte, dass wir bei einfacher Nullhypothese ($\theta = \theta_0$ ist eine solche) und symmetrischer Verteilung (Normalverteilung und t_n -Verteilung) mit Cox-Hinkley (je $\alpha/2$ links und rechts) auch UMPU haben. Deshalb sind die beiden ersten der folgenden KI im Sinne von 5.2.2 optimal.

5.2.3.1 $\mathcal{N}(\mu, \sigma^2)$: KI für μ wenn σ^2 bekannt (optimal)

Gegeben x_1, \dots, x_n , Stichprobe aus $\mathcal{N}(\mu, \sigma^2)$, σ^2 bekannt. Gesucht: KI für μ .

Wir haben diesen einfachen Spezialfall gar nie untersucht; es ist aber klar, dass ein Test nach Cox-Hinkley z.B. mit $\alpha = 5\%$ folgendermassen aussieht: Annahmebereich für $\mu = \mu_0$:

$$\frac{\sqrt{n}}{\sigma} \left| \bar{x} - \mu_0 \right| \leq 1.96.$$

Dies ist also A_{μ_0} . Wir erhalten

$$KI_{\alpha}(\mathbf{x}) = \{\mu | \mathbf{x} \in A_{\mu}\} = \left\{ \mu \mid \frac{\sqrt{n}}{\sigma} \left| \bar{x} - \mu \right| \leq 1.96 \right\} = \left\{ \mu \mid \left| \bar{x} - \mu \right| \leq \frac{1.96\sigma}{\sqrt{n}} \right\}.$$

Damit muss also gelten

$$\mu \in \left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right].$$

$\left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right]$ ist eine Realisation eines 95 % KI für den Mittelwert. Je grösser n , desto kleiner ist das Intervall.

5.2.3.2 $\mathcal{N}(\mu, \sigma^2)$: KI für μ wenn σ^2 unbekannt (optimal)

Gegeben x_1, \dots, x_n , Stichprobe aus $\mathcal{N}(\mu, \sigma^2)$, σ^2 unbekannt. Gesucht: KI für μ . Vergleichen Sie immer 5.2.3.2 mit 5.2.3.1. Von 4.4.3 wissen wir, dass der Annahmehereich für $\mu = \mu_0$ folgendermassen aussieht

$$\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}} \leq qt(0.975, n-1)$$

Dies ist also A_{μ_0} . Wir definieren: $t^* := qt(0.975, n-1)$ und $\hat{\sigma} := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ und erhalten

$$KI_{\alpha}(\mathbf{x}) = \{\mu | \mathbf{x} \in A_{\mu}\} = \left\{ \mu \mid \frac{\sqrt{n}|\bar{x} - \mu|}{\hat{\sigma}} \leq t^* \right\} = \left\{ \mu \mid |\bar{x} - \mu| \leq \frac{t^* \hat{\sigma}}{\sqrt{n}} \right\}.$$

Damit muss also gelten

$$\mu \in \left[\bar{x} - \frac{t^* \hat{\sigma}}{\sqrt{n}}, \bar{x} + \frac{t^* \hat{\sigma}}{\sqrt{n}} \right].$$

$\left[\bar{x} - \frac{t^* \hat{\sigma}}{\sqrt{n}}, \bar{x} + \frac{t^* \hat{\sigma}}{\sqrt{n}} \right]$ ist eine Realisation eines 95 % KI für den Mittelwert. Je grösser n , desto kleiner ist das Intervall. Das Intervall ist tendenziell grösser als das Intervall in 5.2.3.1, weil wir noch die Unsicherheit in der Schätzung von σ haben.

5.2.3.3 $\mathcal{N}(\mu, \sigma^2)$: KI für σ^2 wenn μ unbekannt (nicht optimal aber etabliert)

Gegeben x_1, \dots, x_n , Stichprobe aus $\mathcal{N}(\mu, \sigma^2)$, μ unbekannt. Gesucht: KI für σ^2 . Von 4.4.2 wissen wir, dass der Annahmehbereich für $\sigma^2 = \sigma_0^2$ folgendermassen aussieht (nicht UMPU, aber etabliert)

$$\text{qchisq}(0.025, n-1) \leq \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma_0^2} \leq \text{qchisq}(0.975, n-1).$$

Dies ist also $A_{\sigma_0^2}$. Wir definieren: $a := \text{qchisq}(0.025, n-1)$, $b := \text{qchisq}(0.975, n-1)$, $S^2 := \sum_{i=1}^n (x_i - \bar{x})^2$ und erhalten

$$KI_{\alpha}(\mathbf{x}) = \{\sigma^2 | \mathbf{x} \in A_{\sigma^2}\} = \{\sigma^2 | a \leq \frac{S^2}{\sigma^2} \leq b\}.$$

Damit muss also gelten

$$\sigma^2 \in \left[\frac{S^2}{b}, \frac{S^2}{a} \right].$$

$[S^2/b, S^2/a]$ ist eine Realisation eines 95 % KI für σ^2 . Je grösser n , desto kleiner ist das Intervall auch hier (ziemlich versteckt in den Formeln ...).

5.2.4 Heuristische Konstruktionsmethoden

Das nachfolgende Kochrezept wird in der Praxis sehr häufig angewendet. Wir werden es am Beispiel aus 5.2.3.1 vordemonstrieren.

Kochrezept für ein KI für θ

1. Mit welcher Statistik würde man θ schätzen?
2. brauchen stabilen Ausdruck \Rightarrow zentrieren / normieren
3. kritische Werte (untere/obere 2.5 %) von *bekannter* Verteilung
4. Umformen, bis man es als KI verkaufen kann

5.2.3.1 revisited: $\mathcal{N}(\mu, \sigma^2)$: KI für μ wenn σ^2 bekannt (optimal)

Gegeben x_1, \dots, x_n , Stichprobe aus $\mathcal{N}(\mu, \sigma^2)$, σ^2 bekannt. Gesucht: KI für μ .

1. Mit welcher Statistik würde man μ schätzen?

$$\bar{x}$$

2. brauchen stabilen Ausdruck \Rightarrow zentrieren / normieren

$$\frac{\sqrt{n}}{\sigma} |\bar{X} - \mu|$$

Man beachte: \bar{X} und nicht \bar{x} .

3. kritische Werte (untere/obere 2.5 %) von *bekannter* Verteilung (hier $\mathcal{N}(0, 1)$)

$$P_{\mu} \left[\frac{\sqrt{n}}{\sigma} |\bar{X} - \mu| \leq 1.96 \right] = 0.95$$

4. Umformen, bis man es als KI verkaufen kann

$$\frac{\sqrt{n}}{\sigma} |\bar{x} - \mu| \leq 1.96$$

$$|\bar{x} - \mu| \leq \frac{1.96\sigma}{\sqrt{n}}$$

$$\mu \in \left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right]$$

Es ist dies das gleiche Intervall wie vorher in 5.2.3.1.