

5 Asymptotik und Robustheit

5.1 Konsistenz

- a Das Thema dieses Blockes lautet, wie man die Genauigkeit von aus Beobachtungen abgeleiteten Grössen – vor allem Schätzungen – ermitteln kann. Unter dem Begriff **Asymptotik** fasst man Betrachtungen zur Frage zusammen, wie die Verteilung von solchen Grössen sich verhält, wenn die Anzahl der Beobachtungen immer grösser wird – wenn sie gegen unendlich strebt.

Hier fassen wir zunächst Ergebnisse zusammen, die auch im Buch von Stahel (2007, Kap. 5.8) enthalten sind.

- b In einer Reihe von unabhängigen Versuchen nähert sich die **relative Häufigkeit** R_n des Eintretens eines Ereignisses A immer mehr einem bestimmten Wert, nämlich der Wahrscheinlichkeit $P\langle A \rangle = \pi = \mathcal{E}\langle R_n \rangle$. Anders gesagt: Das Ereignis $\{|R_n - \pi| > \varepsilon\}$, dass die Abweichung der relativen Häufigkeit von der Wahrscheinlichkeit grösser als eine beliebige (kleine) Zahl $\varepsilon > 0$ wird, wird mit wachsendem n immer unwahrscheinlicher,

$$\lim_{n \rightarrow \infty} P\{|R_n - \pi| > \varepsilon\} = 0 .$$

Dies ist die einfachste Variante des so genannten **Gesetzes der grossen Zahl** und geht auf Jakob Bernoulli (publiziert posth. 1713) zurück.

Begründung: Unter n unabhängigen Versuchen ist die Anzahl X derer, in denen ein Ereignis A eintritt, bekanntlich binomial verteilt mit Parameter $\pi = P\langle A \rangle$, $X \sim \mathcal{B}\langle n, \pi \rangle$. Der Erwartungswert der relativen Häufigkeit X/n ist deshalb $\mathcal{E}\langle X/n \rangle = \mathcal{E}\langle X \rangle / n = n\pi/n = \pi$ und die Varianz $\text{var}\langle X/n \rangle = \text{var}\langle X \rangle / n^2 = n\pi(1 - \pi)/n^2 = \pi(1 - \pi)/n$. Schliesslich kommt also für $n \rightarrow \infty$ eine Verteilung mit Erwartungswert π und Varianz 0 heraus – also „mit Sicherheit“ die Zahl π .

- c **Schwaches Gesetz der grossen Zahl.** Diese Idee lässt sich verallgemeinern zur Betrachtung des arithmetischen Mittelwertes von n unabhängigen Beobachtungen einer (fast) beliebigen Zufallsvariablen:

Wenn $X_1, X_2, \dots, X_n \dots$ unabhängige Zufallsvariable mit der gleichen Verteilung sind, gilt

$$P\{|\bar{X}_n - \mu| > \varepsilon\} \xrightarrow{n \rightarrow \infty} 0 \quad \text{für jedes } \varepsilon > 0 ,$$

wobei μ der Erwartungswert der X_i ist. Man nennt diese Eigenschaft **Konvergenz in Wahrscheinlichkeit** oder auch **Konsistenz** von \bar{X} für μ .

Die Überlegung ist genau analog zur vorhergehenden.

- d Die empirische **kumulative Verteilungsfunktion** nähert sich der theoretischen kumulative Verteilungsfunktion, wenn n gross und grösser wird,

$$\widehat{F}_n\langle x \rangle = \frac{1}{n} \text{Anzahl}\langle i : X_i \leq x \rangle \xrightarrow{n \rightarrow \infty} F\langle x \rangle = P\langle X \leq x \rangle .$$

Das folgt aus dem ersten Satz, da es ja um die relative Häufigkeit des Ereignisses $\{X_i \leq x\}$ geht.

- e **Konsistenz der Kennzahlen.** Das macht plausibel, dass alle in Tabelle 5.3.b des Buches zusammengestellten Kennzahlen für Stichproben bei steigendem Stichprobenumfang immer genauer gleich der entsprechenden Kennzahl der Verteilung der einzelnen Beobachtung X_i werden müssen, denn wir haben festgelegt, dass die theoretischen Kennzahlen aus den empirischen entstehen, indem man die relativen Häufigkeiten durch die Wahrscheinlichkeiten ersetzt.
- f **Funktionale.** Die empirischen Kennzahlen kann man ausrechnen, wenn man die empirische Verteilungsfunktion \widehat{F}_n kennt, und ebenso sind die theoretischen Kennzahlen gegeben, wenn die theoretische Verteilungsfunktion F bekannt ist. Die Kennzahlen sind also Funktionen, die jeder Verteilungsfunktion F eine Zahl $T\langle F \rangle$ zuordnen – den theoretischen Verteilungsfunktionen die theoretischen Kennzahlen und den empirischen Verteilungsfunktionen die empirischen Kennzahlen $T\langle \widehat{F}_n \rangle$.
- Weil nun für $n \rightarrow \infty$ $\widehat{F}_n \rightarrow F$ geht, geht auch

$$T\langle \widehat{F}_n \rangle \xrightarrow{n \rightarrow \infty} T\langle F \rangle .$$

* Damit all das sicher gilt, braucht es noch mathematische Voraussetzungen, die bei den gebräuchlichen Verteilungen und Kennzahlen immer erfüllt sind. Die letzte Schlussfolgerung entspricht einer Stetigkeit des Funktionals. Ausserdem muss die Existenz von $T\langle F \rangle$ vorausgesetzt werden. Es gibt Verteilungen, die so langschwänzig sind, dass sie keine Varianz oder sogar keinen Erwartungswert haben, da das Integral, das diese (theoretischen) Kennzahlen definiert, unendlich oder unbestimmt ist.

- g In den Definitionen von Erwartungswert und Varianz von stetigen Zufallsvariablen kommen Integrale vor, die für die empirischen Verteilungen (und für diskrete Zufallsvariable) arithmetischen Mittelwerten entsprechen. Die Mathematiker schreiben allgemein $\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle$ und definieren das so, dass für stetige Zufallsvariable

$$\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle = \int x f\langle x \rangle dx ,$$

für diskrete

$$\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle = \sum_x x \langle X = x \rangle$$

und deshalb für empirische Verteilungsfunktionen

$$\int x d\widehat{F}_n\langle x \rangle = \sum_i x_i \frac{1}{n} = \bar{X}$$

gilt.

- h Wenden wir diese Begriffe auf die Schätzung von Parametern an! Wir betrachten eine parametrische Familie mit einem oder mehreren Parametern $\underline{\theta} = [\theta_1, \dots, \theta_p]$ und entsprechende Stichproben X_1, \dots, X_n .

Wenn eine Stichprobenfunktion $T_k\langle \widehat{F}_n \rangle$ den Parameter θ_k schätzen soll, so wird man fordern, dass 5.1.f für $F = F_{\underline{\theta}}$ und

$$T_k\langle F_{\underline{\theta}} \rangle = \theta_k$$

(für alle Parameterwerte $\underline{\theta}$) gelten soll. Man nennt diese Eigenschaft **Fisher-Konsistenz**.

- i Die Parameter sind bei den drei einfachsten Modellen Binomial-, Poisson- und Normalverteilung gleichzeitig Kennzahlen (Erwartungswert – bis auf einen Faktor n bei der Binomialverteilung – und Varianz, und es ist naheliegend und erst noch optimal, für die Schätzung jeweils die entsprechende(n) empirische(n) Kennzahl(en) zu verwenden.
- j **Das einfache Lokationsmodell.** Das einfachste Problem, an dem die folgenden Überlegungen veranschaulicht werden können, ist die Schätzung des Erwartungswertes μ einer Normalverteilung (oder bald auch einer anderen symmetrischen Verteilung) bei bekanntem Skalenparameter $\sigma = \sigma_0$. Wenn eine Stichprobe

$$X_1, X_2, \dots, X_n, \quad X_i \sim \mathcal{N}\langle \mu, \sigma_0^2 \rangle, \quad \text{unabhängig}$$

gegeben ist, dann ist die naheliegende Schätzung das arithmetische Mittel \bar{X} . Man kann aber auch den Median der Beobachtungen verwenden, denn μ ist auch der Median – der empirische Median ist also auch Fisher-konsistent für μ . Wir werden gleich noch weitere Möglichkeiten kennen lernen.

5.2 Maximum likelihood und M-Schätzungen

- a **Repetition des Prinzips der maximalen Likelihood.** Die Wahrscheinlichkeitsdichte $f_{\underline{\theta}}\langle x \rangle$ eines parametrischen Modells – oder bei diskreten Verteilungen die Wahrscheinlichkeiten $P_{\underline{\theta}}\langle X = x \rangle$ – kann als Funktion der Parameter aufgefasst werden und heisst dann Likelihood; wir schreiben jetzt beide Argumente hinter einander, $f\langle x, \underline{\theta} \rangle$, und das soll auch den diskreten Fall umfassen, $f\langle x, \underline{\theta} \rangle = P_{\underline{\theta}}\langle X = x \rangle$ (In der Mathematik nennt man das eine „Dichte bezüglich dem Zählmass“.)

Die Maximum-Likelihood-Schätzung bestimmt den Parameter θ oder die Parameter $\underline{\theta} = [\theta_1, \dots, \theta_p]$, die $f\langle x, \underline{\theta} \rangle$ maximal machen. Was man beobachtet hat, soll die grösste Wahrscheinlichkeitsdichte erhalten.

- b Wenn man eine Stichprobe von unabhängigen Beobachtungen X_1, \dots, X_n hat, ist die gemeinsame Dichte für alle Beobachtungen wegen der Unabhängigkeit das Produkt der Dichten $f\langle x_i, \underline{\theta} \rangle$. Deshalb maximiert man das Produkt $\prod_i f\langle x_i, \underline{\theta} \rangle$. Einfacher geht es, wenn man stattdessen den Logarithmus des Produkts maximiert; das Resultat (der Wert, für den das Maximum auftritt) ist dasselbe. Der Logarithmus des Produkts ist ja eine Summe. Die Maximum-Likelihood-Schätzung maximiert also

$$L\langle x_1, \dots, x_n; \underline{\theta} \rangle = \sum_i \log\langle f\langle x_i, \underline{\theta} \rangle \rangle$$

über $\underline{\theta}$. Es ist üblich, diese Funktion noch mit -2 zu multiplizieren und dann zu minimieren; man bestimmt also die Minimalstelle von

$$D\langle \underline{\theta} \rangle = -2 \sum_i \log\langle f\langle x_i, \underline{\theta} \rangle \rangle = \sum_i \rho\langle x_i, \underline{\theta} \rangle$$

Die Funktion D heisst **Devianz**. Die ρ -Funktion quantifiziert die „Abweichung“ der Beobachtung x_i vom Modell mit Parameter $\underline{\theta}$. Der Faktor -2 führt dazu, dass für die Normalverteilung mit bekannter Varianz $\rho\langle x, \mu \rangle = ((x_i - \mu)/\sigma)^2$ plus eine Konstante ist. (Die Minimierung der Devianz bezüglich μ führt zu Kleinsten Quadraten.)

- c **Beispiel logistische Verteilung.** Damit es nicht immer die Normalverteilung ist, nehmen wir als Beispiel die logistische Verteilung als Modell. Sie hat die Dichte

$$\frac{1}{(e^{z/2} + e^{-z/2})^2}, \quad z = \frac{x - \mu}{\sigma}$$

Dieses Modell ist, wie die Normalverteilung, dadurch charakterisiert, dass man von einer gegebenen Dichte ausgeht, die symmetrisch um 0 ist, und die parametrische Familie erhält, indem man diese Verteilung mit σ streckt und um μ verschiebt – man nennt das eine **Lokations-Skalen-Familie**. Die Form ist etwas „langschwänziger“ als die Normalverteilung. Die log-likelihood ist $-2 \log \langle e^{z/2} + e^{-z/2} \rangle$.

- d Eine Maximalstelle kann man bestimmen, indem man ableitet und null setzt. Ableitung nach θ_k führt allgemein zu

$$\frac{\partial L}{\partial \theta_k} \langle x_1, \dots, x_n; \underline{\theta} \rangle = \sum_{i=1}^n s_k \langle x_i; \underline{\theta} \rangle,$$

wobei s die so genannten „Likelihood-Scores“ bezeichnet,

$$s_k \langle x; \underline{\theta} \rangle = \frac{\partial}{\partial \theta_k} \log \langle f \langle x_i, \underline{\theta} \rangle \rangle = -\frac{1}{2} \frac{\partial}{\partial \theta_k} \rho \langle x_i, \underline{\theta} \rangle.$$

In den üblichen Fällen wird bei einer Maximalstelle die Ableitung null. Bei mehreren Parametern $\underline{\theta}$ werden alle partiellen Ableitungen null. Wenn man dann die s_k zu einem Vektor \underline{s} zusammenfasst, kann man schreiben

$$\sum_{i=1}^n \underline{s} \langle x_i; \hat{\underline{\theta}} \rangle = 0.$$

Die **Maximum-Likelihood-Schätzung** erhält man dann durch Auflösen dieser (impliziten) Gleichung nach $\hat{\underline{\theta}}$. In einfachen Fällen kann dies explizit, als Formel, geschehen. In anderen Fällen muss man die Lösung numerisch bestimmen. (Dann kann allerdings die direkte numerische Maximierung von L einfacher sein.)

- e Im Fall der **logistischen Verteilung** wird, da $z = (x - \mu)/\sigma$ und damit $\partial z / \partial \mu = -1/\sigma$ und $\partial z / \partial \sigma = -(x - \mu)/\sigma^2 = -z/\sigma$ ist,

$$\begin{aligned} s_\mu \langle x; \underline{\theta} \rangle &= \frac{1}{\sigma} \frac{e^{z/2} - e^{-z/2}}{e^{z/2} + e^{-z/2}} \\ s_\sigma \langle x; \underline{\theta} \rangle &= z s_\mu \langle x; \underline{\theta} \rangle. \end{aligned}$$

- f Zu jedem Modell gehört also eine Maximum-Likelihood-Schätzung, wie sie gerade allgemein definiert wurde. Nun wollen wir uns von dieser strikten Zuordnung lossagen und nehmen uns die Freiheit, Schätzfunktionen anders, aber immer noch analog zu einer Maximum-Likelihood-Schätzung festzulegen, entweder durch eine Funktion ρ in 5.2.b oder durch eine Funktion $\underline{\psi}$ (statt \underline{s}) in 5.2.d,

$$\begin{aligned} \hat{\underline{\theta}} &= \operatorname{argmin}_{\underline{\theta}} \sum_{i=1}^n \rho \langle X_i, \underline{\theta} \rangle \quad \text{oder} \\ \hat{\underline{\theta}} &= \text{Lösung von } \sum_{i=1}^n \underline{\psi} \langle X_i, \underline{\theta} \rangle = 0 \end{aligned}$$

(argmin heisst Minimalstelle). Die Funktionen ρ und $\underline{\psi}$ müssen nicht unbedingt von einem Modell hergeleitet sein; ρ soll einfach auf sinnvolle Weise messen, wie schlecht eine Beobachtung x zum Parameter (-Vektor) $\underline{\theta}$ passt; $\underline{\psi}$ soll generell die Eigenschaften zeigen,

die partielle Ableitungen einer solchen Funktion haben. Da die Minimierung des ersten Ausdrucks fast immer mit dem null Setzen der partiellen Ableitungen äquivalent ist, kann man jede Schätzung gemäss der ersten Definition auch als eine mit der zweiten Definition schreiben. Umgekehrt gibt es hingegen auch sinnvolle $\underline{\psi}$ -Funktionen, die sich nicht als partielle Ableitungen einer ρ -Funktion schreiben lassen. Die zweite Definition ist also noch allgemeiner als die erste. Wir werden von jetzt an nur mit der zweiten Definition arbeiten, vor allem, weil wichtige Eigenschaften der Schätzung direkt mit $\underline{\psi}$ zusammenhängen.

Man könnte solche Schätzungen verallgemeinerte Maximum-Likelihood-Schätzungen nennen, sie heissen aber einfach **M-Schätzungen**.

Wir können beispielsweise die scores-Funktion der logistischen Verteilung benutzen, auch wenn wir eigentlich hoffen oder glauben, dass die Beobachtungen der Normalverteilung folgen – und wir werden sehen, dass so etwas gute Gründe hat.

- g **M-Schätzungen als Funktionale.** In der Definition der M-Schätzungen treten Summen auf, die man gerade so gut durch Mittelwerte ersetzen könnte. Mittelwerte entsprechen Erwartungswerten, und so kann man jede M-Schätzung zu einem Funktional machen, indem man sie definiert als

$$T_{\rho}\langle F \rangle = \operatorname{argmin}_{\underline{\theta}} \int \rho\langle x, \underline{\theta} \rangle dF\langle x \rangle \quad \text{oder}$$

$$T_{\psi}\langle F \rangle = \text{Lösung von } \int \underline{\psi}\langle x, \underline{\theta} \rangle dF\langle x \rangle = 0 .$$

Die Schätzungen erhält man, indem man als F die empirische Verteilungs-Funktion \widehat{F}_n einsetzt. Da Mittelwerte über $\rho\langle X_i, \underline{\theta} \rangle$ oder $\underline{\psi}\langle X_i, \underline{\theta} \rangle$ – Integrale über \widehat{F}_n – für $n \rightarrow \infty$ in Integrale über die theoretische Verteilung F der Beobachtungen übergehen, ist es naheliegend, dass die Minimalstellen respektive die Nullstellen, die die Schätzungen definieren, nach $T\langle F \rangle$ gehen, also 5.1.f gilt. M-Schätzungen sind konsistent.

Wenn die M-Schätzung einen Parameter θ schätzen soll, dann muss $T\langle F_{\underline{\theta}} \rangle = \underline{\theta}$ sein. Für $\underline{\psi}$ heisst das, dass

$$\int \underline{\psi}\langle x, \underline{\theta} \rangle dF_{\underline{\theta}}\langle x \rangle = 0$$

sein muss für alle $\underline{\theta}$. Die Schätzung ist dann Fisher-konsistent.

- h **Beispiel Lokationsmodell und Huber-Schätzung.** Im Lokationsmodell ist die Dichte, wenn wir $\sigma_0 = 1$ setzen,

$$f\langle x, \mu \rangle = f\langle x - \mu, 0 \rangle , \quad f\langle z, 0 \rangle = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} .$$

Die log-likelihood ist $-\frac{1}{2} \log\langle 2\pi \rangle - (x - \mu)^2/2$ und die Scores werden $s\langle x, \mu \rangle = x - \mu$. Einfacher geht es kaum. Die Maximum-Likelihood-Schätzung wird deshalb die Lösung von $\sum_i (x_i - \widehat{\mu}) = 0$ oder $\widehat{\mu} = \frac{1}{n} \sum_i x_i$: das arithmetische Mittel, wie bekannt. Man sieht, dass das arithmetische Mittel eine M-Schätzung ist, gegeben durch $\psi\langle x, \mu \rangle = x - \mu$.

Der Median lässt sich ebenfalls als M-Schätzung schreiben: Man setzt

$$\psi\langle x, \mu \rangle = \begin{cases} -1 & x - \mu < 0 \\ 1 & x - \mu > 0 \end{cases} .$$

Als Kompromiss zwischen diesen beiden Schätzungen hat Prof. Peter Huber die M-Schätzungen

mit einer ψ -Funktion der Form

$$\psi(x, \mu) = \begin{cases} x - \mu & \text{für } |x - \mu| \leq k \\ -k & \text{für } x - \mu < -k \\ k & \text{für } x - \mu > k \end{cases}$$

eingeführt; k ist eine wählbare Konstante (tuning constant). Diese „Huber-Schätzungen“ spielen in der Theorie der robusten Schätzungen (siehe 5.6) eine zentrale Rolle.

Figur 5.2.h zeigt die besprochenen und eine weitere ψ -Funktion im Vergleich.

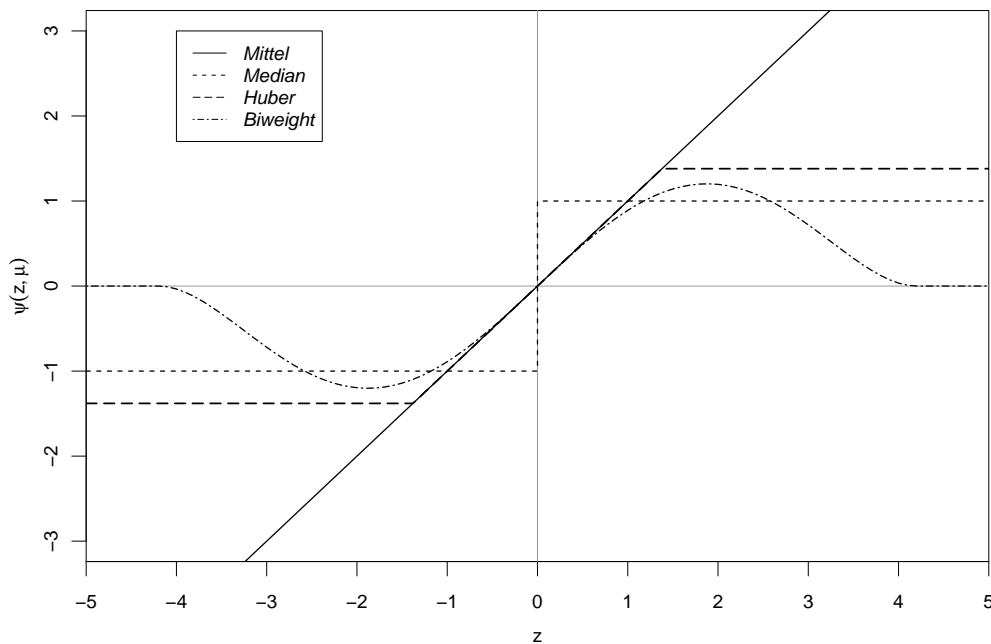


Abbildung 5.2.h: ψ -Funktionen für gebräuchliche M-Schätzungen

- i Kehren wir zu den Maximum-Likelihood-Schätzungen zurück, die ja spezielle M-Schätzungen mit $\underline{\psi} = \underline{s}$ sind! Man kann leicht zeigen, dass allgemein

$$\int \underline{s}\langle x, \underline{\theta} \rangle f\langle x, \underline{\theta} \rangle dx = 0$$

gilt. Das heisst, dass Maximum-Likelihood-Schätzungen immer Fisher-konsistente Schätzungen für die Parameter der Verteilungsfamilie sind.

Beweis: Leitet man $\int f\langle x, \underline{\theta} \rangle dx = 1$ nach θ ab, so erhält man $\int \frac{\partial}{\partial \theta} f\langle x, \underline{\theta} \rangle dx = 0$. Wegen der Definition von \underline{s} ist $\frac{\partial}{\partial \theta} f\langle x, \underline{\theta} \rangle = \underline{s}\langle x, \underline{\theta} \rangle f\langle x, \underline{\theta} \rangle$, und Einsetzen ergibt die Gleichung.

5.3 Einflussfunktion

- a Kehren wir von den Betrachtungen von Grenzwerten für grosse Stichproben zu den einzelnen Beobachtungen zurück. Wir fragen, welchen Einfluss eine einzelne Beobachtung auf einen Schätzwert hat.
- b **Empirische Einflussfunktion.** Wir gehen von einer beobachteten Stichprobe aus – konkret von 10 der 100 Werten des **Beispiels der Kükengewichte** (siehe 2.4.a in Stahel (2007)), nämlich

$$107 \quad 108 \quad 111 \quad 101 \quad 97 \quad 113 \quad 109 \quad 105 \quad 116 \quad 122 .$$

Wir „probieren aus“, wie eine zusätzliche Beobachtung mit dem x-beliebigen Wert x_0 den Wert des arithmetischen Mittels verändern würde. Das ist einfach auszurechnen: Der Mittelwert aller Beobachtungen ist

$$\begin{aligned} (107 + 108 + \dots + 122 + x_0)/11 &= \frac{n\bar{x} + x_0}{n+1} = \frac{(n+1)\bar{x} - \bar{x}}{n+1} + \frac{x_0}{n+1} = \bar{x} + \frac{1}{n+1}(x_0 - \bar{x}) \\ &= 108.9 + (x_0 - 108.9)/11 . \end{aligned}$$

Die gleiche Frage können wir auch für andere Schätzungen stellen. Der Median wird von $(108 + 109)/2 = 108.5$ zu 108, falls $x_0 \leq 108$ ist, und zu 109, falls $x_0 \geq 109$ ausfällt.

In Figur 5.3.b werden diese Schätzwerte als Funktion der zusätzlichen Beobachtung x_0 dargestellt, zusammen mit dem 10%-gestutzten Mittelwert (hier: kleinste und grösste Beobachtung weglassen und Mittel der verbleibenden ausrechnen).

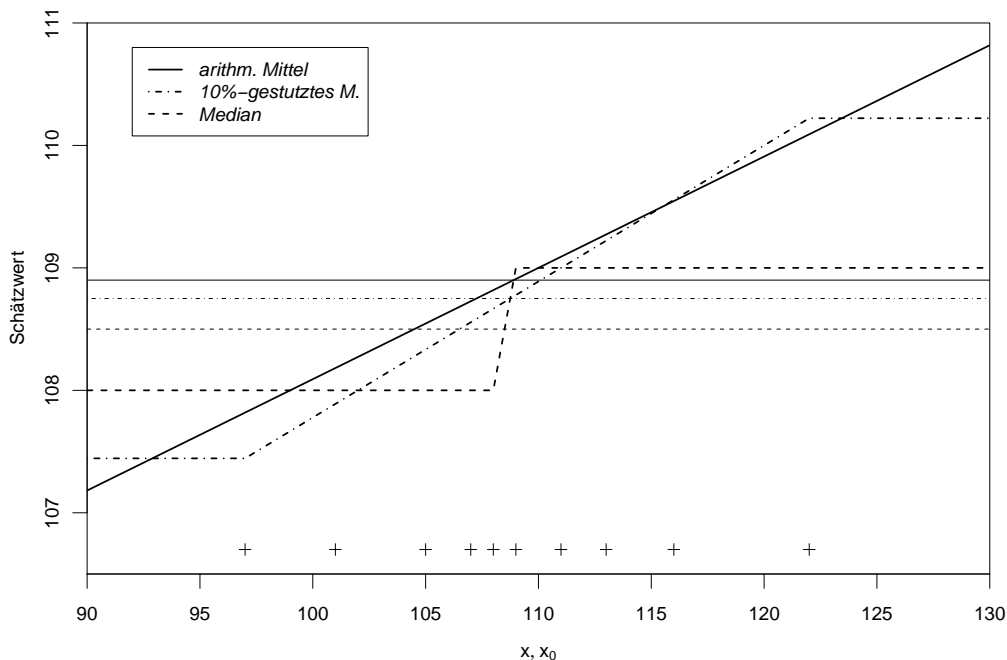


Abbildung 5.3.b: Effekt einer zusätzlichen Beobachtung mit Wert x_0 auf die Schätzungen von μ . Die gegebenen Beobachtungen sind durch die Kreuze (+) repräsentiert. Der Wert der Schätzungen für die ursprüngliche Stichprobe ist jeweils als horizontale Linie eingezeichnet.

- c Die Veränderung durch den Zusatzwert x_0 hängt natürlich von der Stichprobengröße n ab. Es ist höchst plausibel und auch für das gewöhnliche und das gestutzte Mittel leicht aus den Formeln herauszulesen, dass der Einfluss jeder einzelnen Beobachtung proportional zu $1/n$ ist. Um diesen trivialen Effekt los zu werden, multipliziert man die Veränderung mit n und definiert allgemein

$$SC\langle x_0; T, x_1, \dots, x_n \rangle = n (T\langle x_1, \dots, x_n, x_0 \rangle - T\langle x_1, \dots, x_n \rangle)$$

als die **empirische Einflussfunktion** oder **Sensitivity Curve** der Schätzung (für die gegebene Stichprobe).

- d **Gross Error Model.** Was geschieht, wenn nun n gegen ∞ geht? Die Stichprobe, genauer: die empirische Verteilungsfunktion, wird ja zur theoretischen Verteilungsfunktion F . Eine zusätzliche Beobachtung verschwindet dann natürlich in den „unendlich vielen“ der Stichprobe. Ersetzen wir die eine Beobachtung durch einen (kleinen) Anteil ε von zusätzlichen Beobachtungen, die alle gleich x_0 sind, und untersuchen ihren Effekt. Sie führen zu einer kleinen Stufe der Höhe ε in der Verteilungsfunktion. Wir erhalten als kumulative Verteilungsfunktion der „Misch-Verteilung“ von einem Anteil $1 - \varepsilon$ von „gewünschten“ Beobachtungen, die dem Modell F folgen, und einem Anteil ε von „unerwünschten“ Beobachtungen, die alle gleich x_0 sind,

$$(1 - \varepsilon) F\langle \cdot \rangle + \varepsilon \Delta_{x_0}\langle \cdot \rangle$$

wobei Δ_{x_0} die „Stufen-Funktion“ ist, die bei x von 0 auf 1 springt, $\Delta_{x_0}\langle x \rangle = 0$ für $x \leq x_0$ und $= 1$ für $x > x_0$.

Je nach dem Wert x_0 führt ein „grober Fehler“ zu einem Ausreisser oder einer durchaus plausiblen Beobachtung (Figur 5.3.d).

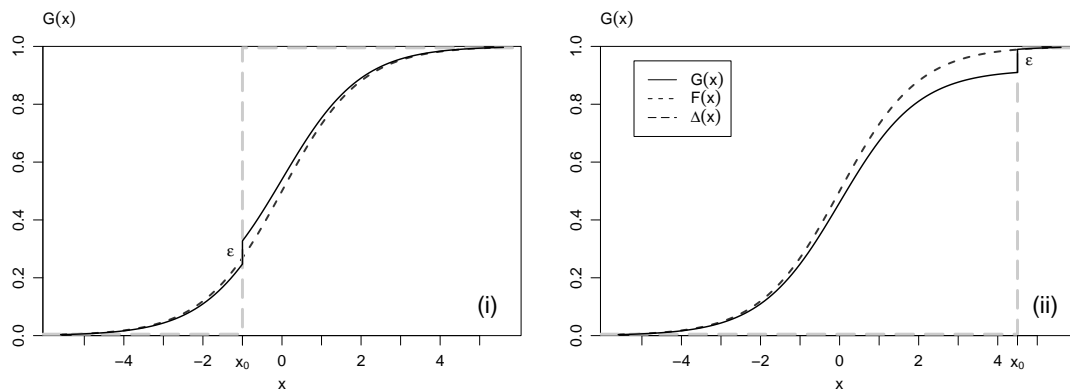


Abbildung 5.3.d: Modell für einen „groben Fehler“ bei x_0 . (i) x_0 liegt im Bereich der erwarteten guten“ Beobachtungen, (ii) der grobe Fehler bildet einen Ausreisser.

Allgemeiner lässt man für die groben Fehler nicht nur einen Wert, sondern eine beliebige andere Verteilung zu und erhält dann die Mischverteilung mit

$$G\langle x \rangle = (1 - \varepsilon) F\langle x \rangle + \varepsilon H\langle x \rangle ,$$

das so genannte Gross Error Model.

- e **Einflussfunktion.** Das gerade eingeführte Modell (in der einfacheren Form) dient dazu, die empirische Einflussfunktion von den Daten unabhängig zu machen. Die (asymptotische) Einflussfunktion ist definiert als

$$\text{IF}\langle x; T, F \rangle = \lim_{\varepsilon \rightarrow 0} \frac{T\langle (1 - \varepsilon)F + \varepsilon\Delta_x \rangle - T\langle F \rangle}{\varepsilon} .$$

- f Haben Sie mit Grenzwerten keine Übung? Einige sind einfach auszurechnen.

Das arithmetische Mittel \bar{X} ist für theoretische Verteilungen ja der Erwartungswert, und der erfüllt

$$\mathcal{E}\langle (1 - \varepsilon)F + \varepsilon\Delta_x \rangle = (1 - \varepsilon)\mathcal{E}\langle F \rangle + \varepsilon\mathcal{E}\langle \Delta_x \rangle = (1 - \varepsilon)\mathcal{E}\langle F \rangle + \varepsilon x .$$

Ziehen wir $\mathcal{E}\langle F \rangle$ ab und dividieren durch ε , so erhalten wir $x - \mathcal{E}\langle F \rangle$, was nicht mehr von ε abhängt. Also gibt es keinen Grenzwert zu bilden und

$$\text{IF}\langle x; \bar{X}, F \rangle = x - \mathcal{E}\langle F \rangle .$$

- g* Für den Median ist die Betrachtung auch nicht schwierig. Es ist ja $\text{med}\langle F \rangle = F^{-1}\langle 0.5 \rangle$. Falls nun x grösser als der Median von F ist (genauer, als $t_\varepsilon = F^{-1}\langle 0.5/(1 - \varepsilon) \rangle$), dann ist $\text{med}\langle (1 - \varepsilon)F + \varepsilon\Delta_x \rangle = t_\varepsilon$; für kleine x ergibt sich dementsprechend

$$\text{med}\langle (1 - \varepsilon)F + \varepsilon\Delta_x \rangle = F^{-1}\langle 1 - 0.5/(1 - \varepsilon) \rangle$$

Wenn wir von diesen beiden Ausdrücken $\text{med}\langle F \rangle$ abzählen, die Differenz durch ε dividieren und die Grenzwerte für $\varepsilon \rightarrow 0$ berechnen, dann entspricht das genau der Bestimmung der Ableitung nach ε für $\varepsilon = 0$. Die Ableitung gibt es nur, wenn die Verteilung (in der Nähe des Medians) eine Dichte hat, also für stetige Zufallsvariable. Dann gilt

$$\frac{d}{d\varepsilon} F^{-1}\langle 0.5/(1 - \varepsilon) \rangle = (f\langle F^{-1}\langle 0.5/(1 - \varepsilon) \rangle \rangle)^{-1} \cdot (-0.5) \cdot \frac{-1}{(1 - \varepsilon)^2}$$

was für $\varepsilon = 0$ zu $1/(2f\langle \mu \rangle)$ wird, wobei $\mu = F^{-1}\langle 0.5 \rangle$ der Median ist. Für kleine x kommt der gleiche Wert heraus, mit negativem Vorzeichen. Zwischen den Grenzen für „klein“ und „gross“ liegt ein Bereich, für den wir keine Rechnung gemacht haben – aber dieser Bereich verschwindet für $\varepsilon \rightarrow 0$. Damit wird das Resultat

$$\text{IF}\langle x; \bar{X}, F \rangle = \begin{cases} -1/(2f\langle \mu \rangle) & \text{für } x < \text{med}\langle F \rangle \\ 1/(2f\langle \mu \rangle) & \text{für } x > \text{med}\langle F \rangle \end{cases}$$

Diese Einflussfunktion ist also unstetig beim Median; sie springt von $-1/(2f\langle \mu \rangle)$ auf $1/(2f\langle \mu \rangle)$.

- h **Einflussfunktion für M-Schätzungen.** Es sei θ der (eindimensionale) Parameter eines parametrischen Modells und T_ψ die durch $\psi\langle x, \theta \rangle$ festgelegte Schätzung. Dann gilt

$$\text{IF}\langle x; T, F \rangle = \frac{1}{c}\psi\langle x, \theta \rangle \quad \text{mit} \quad c = - \int \frac{\partial}{\partial \theta} \psi\langle x, \theta \rangle f\langle x, \theta \rangle dx .$$

Die Einflussfunktion von M-Schätzungen ist also proportional zu ψ . Wenn T den Parameter θ schätzt, also $\int \underline{s}\langle x, \underline{\theta} \rangle f\langle x, \underline{\theta} \rangle dx = 0$ gilt (siehe 5.2.g), dann kann man die Konstante c anders schreiben:

$$c = \int \psi\langle x, \underline{\theta} \rangle s\langle x, \underline{\theta} \rangle f\langle x, \underline{\theta} \rangle dx .$$

Das ist oft einfacher zu rechnen.

Spezialfall: Wenn man die Verteilung einer Maximum-Likelihood-Schätzung untersucht für die „Ideal-Annahme“, dass die X_i von der entsprechenden Verteilung kommen, dann liefert das

$$c = - \int \frac{\partial}{\partial \theta} s\langle X, \theta \rangle f\langle x, \underline{\theta} \rangle dx = \int s\langle X, \theta \rangle^2 f\langle x, \underline{\theta} \rangle dx .$$

* Zum Beweis der allgemeinen Formel: In der Definition von $T\langle G \rangle$ für die Verteilung $G = (1 - \varepsilon)F + \varepsilon\Delta_x$ benötigen wir das Integral

$$\int \psi\langle x, T\langle G \rangle \rangle dG\langle x \rangle = (1 - \varepsilon) \int \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle + \varepsilon\psi\langle x, T\langle G \rangle \rangle .$$

Wir linearisieren

$$\psi\langle x, T\langle G \rangle \rangle \approx \psi\langle x, T\langle F \rangle \rangle + \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle (T\langle G \rangle - T\langle F \rangle)$$

und setzen in Integral ein,

$$\int \psi\langle x, T\langle G \rangle \rangle dF\langle x \rangle \approx \int \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle + (T\langle G \rangle - T\langle F \rangle) \int \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle$$

Das erste Integral ist null dank der Definition von $T\langle F \rangle$. Es bleibt

$$(1 - \varepsilon)(T\langle G \rangle - T\langle F \rangle) \int \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle + \varepsilon \left(\psi\langle x, T\langle F \rangle \rangle + (T\langle G \rangle - T\langle F \rangle) \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle \right) = 0$$

nach $T\langle G \rangle - T\langle F \rangle$ aufzulösen,

$$T\langle G \rangle - T\langle F \rangle \approx - \frac{\varepsilon (\psi\langle x, T\langle F \rangle \rangle + (T\langle G \rangle - T\langle F \rangle) \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle)}{(1 - \varepsilon) \int \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle} .$$

Dividiert man nun durch ε und lässt ε gegen 0 gehen, dann erhält man die Einflussfunktion.

- i **Linearisierung.** Es ist naheliegend, anzunehmen, dass sich der Wert einer Schätzung für eine Stichprobe – genauer: seine Abweichung vom Idealwert $T\langle F \rangle$ – näherungsweise als Mittelwert der Einflüsse aller Beobachtungen schreiben lässt,

$$T\langle \widehat{F}_n \rangle \approx T\langle F \rangle + \frac{1}{n} \sum_{i=1}^n \text{IF}\langle X_i; T, F \rangle .$$

* Es lässt sich sogar allgemeiner schreiben

$$T\langle G \rangle \approx T\langle F \rangle + \int \text{IF}\langle x; T, F \rangle d(G - F) .$$

Das erinnert stark an die Linearisierung einer gewöhnlichen Funktion, wobei die Einflussfunktion die Rolle der Ableitung übernimmt. Diese Betrachtung geht auf Richard von Mises zurück. Funktionale T , die dies erfüllen, heißen von Mises-Funktionale, und sie umfassen alle Stichprobenfunktionen, die wir betrachten.

5.4 Asymptotische Verteilung

- a **Zentraler Grenzwertsatz.** Wenn der Stichprobenumfang immer grösser wird, dann wird ein arithmetischer Mittelwert immer genauer gleich dem entsprechenden Erwartungswert. Abb. ... zeigt auch, dass die die Form der Verteilung der noch verbleibenden Abweichungen immer mehr der Form ähnelt, die man bestens kennt. Genauer:

Die Verteilung des standardisierten Mittelwertes

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

nähert sich für wachsendes n immer mehr einer Standard-Normalverteilung an,

$$P\langle Z_n \leq z \rangle \xrightarrow{n \rightarrow \infty} \Phi\langle z \rangle$$

(Φ bezeichnet die kumulative Verteilungsfunktion der Standard-Normalverteilung.) Dies gilt, wenn die X_i unabhängig und gleich verteilt sind und ihre Varianz endlich ist.

Zum Beweis dieses Satzes führen mehrere Wege. Man braucht aber weitere Begriffe oder ziemlich viel Zeit dafür.

- b Man kann den Zentralen Grenzwertsatz auch so schreiben:

$$\begin{aligned} \bar{X}_n &\approx\approx \mathcal{N}\langle \mu, \sigma^2/n \rangle \\ S_n &\approx\approx \mathcal{N}\langle \mu, n\sigma^2 \rangle \end{aligned}$$

($\approx\approx$: „ungefähr verteilt wie ...“, also insgesamt „ungefähr normalverteilt mit Erwartungswert μ und Varianz σ^2/n respektive $n\sigma^2$)

- c **Zentraler Grenzwertsatz für Funktionale.** Aus der Linearisierung in 5.3.i und dem Zentralen Grenzwertsatz ergibt sich

$$T\langle X_1, X_i, \dots, X_n \rangle \approx\approx \mathcal{N}\langle T\langle F \rangle, v/n \rangle \quad v = \text{var} \langle \text{IF}\langle X; F \rangle \rangle$$

Es gilt immer $\mathcal{E} \langle \text{IF}\langle X; F \rangle \rangle = 0$. Deshalb kann man auch schreiben

$$v = \mathcal{E} \langle \text{IF}\langle X; F \rangle^2 \rangle .$$

- d **Asymptotische Varianz für M-Schätzungen.** Aus 5.3.h erhält man

$$v = \frac{1}{c^2} \int \psi\langle x, \theta \rangle^2 dF\langle x \rangle .$$

Für die Maximum-Likelihood-Schätzung ist das Integral gleich dem c , und man erhält

$$v = 1/c, \quad c = \int s\langle x, \theta \rangle^2 dF_\theta\langle x \rangle .$$

Dieses Integral, der Erwartungswert der quadrierten Scores, heisst **Fisher-Information**.

- e **Beispiel Huber-Schätzung.** Für die Huber-Schätzung wird die asymptotischen Varianz bei der Standard-Normalverteilung gleich

$$v = \frac{\int \psi(z, 0)^2 d\Phi(z)}{(\int \psi'(z, 0) d\Phi(z))^2}$$

Dank der einfachen Form von ψ wird der Nenner zum Quadrat von $P(|Z| < k) = 2\Phi(k) - 1$. (Den Zähler kann man durch die kumulative χ^2 -Verteilungsfunktion ausdrücken.)

Figur 5.4.e zeigt die asymptotische Varianz als Funktion der Wahlkonstanten k . Für $k \rightarrow \infty$ geht die Kurve gegen 1, denn die Huber-Schätzung wird dann zum arithmetischen Mittel.

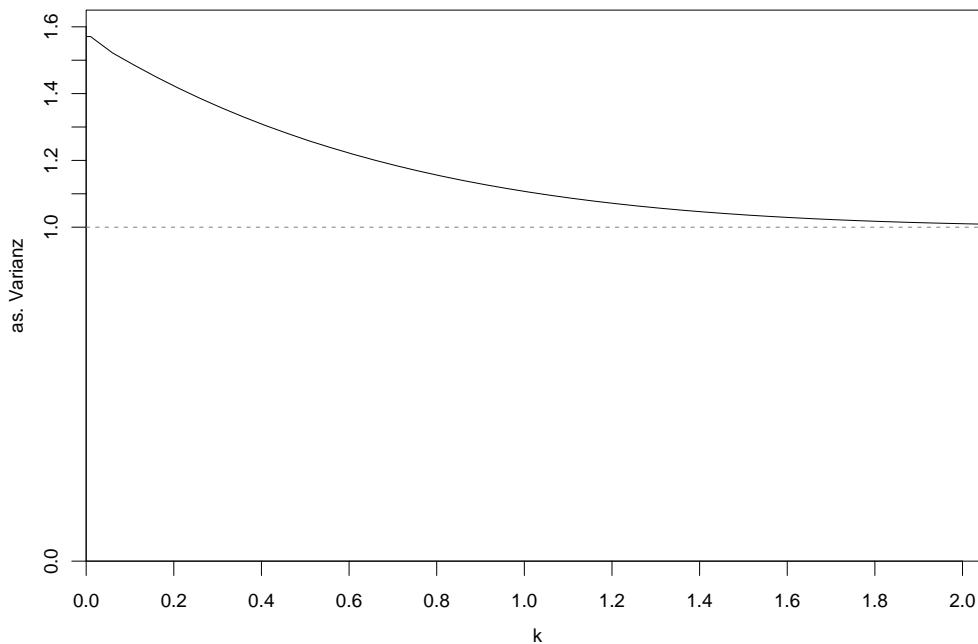


Abbildung 5.4.e: Asymptotische Varianz der Huber-Schätzung in Abhängigkeit vom Wahlparameter k

- f **Beispiel logistische Verteilung.** Die asymptotische Varianz für die Maximum-Likelihood-Schätzung wird für die Verteilung mit $\mu = 0$ und $\sigma = 1$ gleich

$$\int \frac{(e^{z/2} - e^{-z/2})^2}{(e^{z/2} + e^{-z/2})^4} dz = 0.333$$

Die „logistische Standard-Verteilung“ ($\mu = 0, \sigma = 1$) ist nicht direkt mit der Standard-Normalverteilung zu vergleichen, denn sie hat eine grössere Streuung. Ihre Varianz ist $\pi^2/3 = 3.3$. Da die Varianz sozusagen zur Normalverteilung gehört, ist es wohl angebrachter, die asymptotische Varianz der Maximum-Likelihood-Schätzung für die Lokation zu vergleichen. Das ist für die Normalverteilung das arithmetische Mittel mit asymptotischer Varianz 1 bei der Standard-Normalverteilung. Für die logistische Verteilung mit $\sigma = 1$ ist diese Varianz gleich 0.333 und deshalb gleich 1 für den Parameterwert $\sigma = 1.732$.

- g **Maximum likelihood als asymptotisch beste Schätzung.** Man kann beweisen, dass die Maximum-Likelihood-Schätzung unter allen Fisher-konsistenten Schätzungen für den Parameter θ die kleinste asymptotische Varianz hat. Sie ist also die beste Schätzung – wenigstens für (unendlich) grosse Stichproben.
- h **Tests und Vertrauensintervalle.** Wenn die Verteilung einer Schätzung bekannt ist, kann man damit ja ohne Weiteres einen Test für Nullhypothesen über den geschätzten Parameter und ein Vertrauensintervall bestimmen. Da die Verteilung normal ist, erhält man einfach
- als standardisierte Testgrösse $T = (\hat{\theta} - \theta_0) / \sqrt{v/n}$; sie ist näherungsweise standard-normalverteilt.
 - als Vertrauensintervall für θ : $\hat{\theta} \pm 1.96 * \sqrt{v/n}$.
- i **Quadratsummen.** Etliche Teststatistiken sind auf Summen von quadrierten Zufallsvariablen aufgebaut. Ein Beispiel dafür bildet die Teststatistik eines Chiquadrat-Anpassungstests oder eines Tests für Unabhängigkeit in Kontingenztafeln. Beide Grössen sind nach dem „Rezept“

$$T = \sum_k (N_k - \mu_k)^2 / \mu_k, \quad \mu_k = \mathcal{E}\langle N_k \rangle$$

aufgebaut. Da es sich um eine Summe handelt, muss der Zentrale Grenzwertsatz gelten. Das tut er auch, wenn die Anzahl der Terme in der Summe „gross“ ist. Aber man braucht ihn nicht, da eine bessere Approximation durch die Chiquadrat-Verteilung gegeben ist. (Da besteht kein Widerspruch, da die Chiquadrat-Verteilung für eine grosse Anzahl Freiheitsgrade einer Normalverteilung gleicht.)

Wieso eine Chiquadrat-Verteilung? Erinnern wir uns an ihre Definition. Danach ist die χ^2 -Verteilung die Verteilung einer Summe von unabhängigen, quadrierten standard-normalverteilten Zufallsvariablen. Da jeder Summand in der obigen Summe in grober Näherung standard-normalverteilt ist, bildet die Chiquadrat-Verteilung eine gute Approximation der Verteilung der Summe.

- j Likelihood ratio test.

5.5 Likelihood Ratio Tests

Dieser Abschnitt ist für das Skript noch nicht formuliert. Deshalb wird hier der entsprechende Ausschnitt aus den englischen Folien eingefügt. The plausibility of a model in the light of data is measured by the likelihood.

A null hypothesis usually restricts a parameter to a specific value (or one side of a given value, for one-sided case).

The restriction deteriorates the “fit” of the data to the model. The likelihood decreases. If it decreases too much, the null hypothesis must be rejected.

—> Test statistic:

- **likelihood ratio**, or
- log likelihood difference, or
- **deviance** – difference of deviance values between “full model” (free parameter) and “reduced model” (parameter fixed at “null value”)

- b Example: simple regression, scale known. Log likelihood: $c - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_1 x_i - \beta_0)^2$
Maximum likelihood = Least Squares.

Null hypothesis $\beta_1 = 0$, β_0 unspecified. \rightarrow **Log likelihood difference**

$$\begin{aligned} & c - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_1 x_i - \beta_0)^2 - \left(c - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_1 x_i - \beta_0)^2 \right) \\ & = \frac{1}{2\sigma^2} \left(\sum_i (y_i - \beta_1 x_i - \beta_0)^2 - \sum_i (y_i - \beta_0)^2 \right) \end{aligned}$$

= difference of Sums of Squares (total minus residual)
= **Sums of Squares of Model** ... divided by $2\sigma^2$.

σ^2 unknown \rightarrow estimate from residuals! multiply by 2 \rightarrow **Difference of deviances**
 \rightarrow **F-Test**.

(To be precise, σ is estimated under the alternative, not under the null hypothesis...)

Also applicable for multiple regression, more than one coefficient to be tested.

- c Same properties for deviance differences – asymptotically – in general (under conditions):
Under the null hypothesis, **the deviance difference** (= twice the log likelihood ratio) **is distributed asymptotically** $\sim \chi_{df}^2$. Degrees of freedom df = number of parameters that are fixed by null hypothesis.

Only applies to “nested” models: The reduced model is obtained by restricting the full model.

5.6 Robuste Schätzungen

- a Die Einflussfunktion zeigt, wie gross der Einfluss einer einzelnen Beobachtung auf eine Schätzung ist. Auf das arithmetische Mittel wird ein einzelner Wert, wenn er weit von den übrigen Beobachtungen weg liegt – ein Ausreisser – einen sehr grossen Einfluss haben. Die Einflussfunktion ist ja unbegrenzt! Besser sieht es für die beiden anderen Schätzungen in Figur 5.3.b und auch für die Maximum-Likelihood-Schätzung der logistischen Verteilung aus: Ihre Einflussfunktionen sind begrenzt.
- b **Gross error sensitivity.** Es liegt nach dieser Bemerkung nahe, eine begrenzte Einflussfunktion zu verlangen, wenn man sich gegen einen zu starken Einfluss von Ausreissern schützen will, und man kann den maximalen Wert der Einflussfunktion als Mass für die Robustheit verwenden.

Die Gross error sensitivity ist definiert als

$$\gamma^* \langle T, F \rangle = \sup_x \langle |IF \langle x; T, F \rangle| \rangle .$$

(sup heisst supremum und ist der mathematisch präzise Ausdruck für das Maximum.)

- c Beispiele:

- Für den Median ist $|IF \langle x; T, F \rangle| = 1/(2f \langle \text{med} \langle F \rangle \rangle)$ für alle x , also ist auch $\gamma^* \langle T, F \rangle$ so gross.
- Für das arithmetische Mittel ist IF unbegrenzt und deshalb $\gamma^* \langle \bar{X}, F \rangle = \infty$.

- d **Maximaler Bias.** Allerdings ist die Einflussfunktion, wie andere „Ableitungen“ auch, nur für kleine Abweichungen von der Verteilung F , um die herum linearisiert wird, massgebend. Zuverlässiger ist es, die Funktion selbst zu untersuchen.

Wir wollen immer noch annehmen, dass die Modell-Verteilung F näherungsweise gilt. Das drücken wir jetzt so aus, dass die wahre, unbekannte Verteilung G eine Gross Error-Verteilung ist, die sich also schreiben lässt als $(1 - \tilde{\varepsilon})F(\cdot) + \tilde{\varepsilon}H(\cdot)$ mit beliebigem H , aber einem $\tilde{\varepsilon}$, das kleiner oder gleich einer Schranke ε ist. Alle solchen Verteilung bilden die so genannte „**Gross Error-Umgebung**“ $U(F, \varepsilon)$ von F mit „Radius“ ε . (Im mathematischen Sinn ist das keine Umgebung!)

Für die Robustheit fragen wir nach der schlimmsten „Verfälschung“ (Bias), die die Statistik T erfahren kann,

$$b(\varepsilon; T, F) = \sup_{G \in U(F, \varepsilon)} \langle G(T) \rangle .$$

Dieses Mass ist abhängig vom „Radius“ ε und oft schwierig zu berechnen, aber es ist ein befriedigenderes Mass für die Robustheit als die Gross Error Sensitivity, da es auch für Statistiken anwendbar ist, die nicht oder nicht gut linearisierbar sind.

- e Wenn die Gross Error Sensitivity unendlich ist, ist es fast immer auch der maximale Bias, für jedes $\varepsilon > 0$. Das bedeutet, dass ein einzelner Ausreisser, wenn er nur genügend weit weg ist, die Statistik T beliebig weit vom „Sollwert“ $T(F)$ weg bringen kann. Wenn das der Fall ist, spricht man vom „Zusammenbruch“, englisch **breakdown** von T . Das arithmetische Mittel kann man schon mit einem beliebig kleinen „Ausreisser-Anteil“ ε zum Zusammenbruch bringen; für andere Schätzer braucht das einen gewissen Mindestanteil. Diese Überlegungen führen zur folgenden Definition.

Der **Bruchpunkt (breakdown point)** $\varepsilon^*(T, F)$ (bezüglich des Gross Error-Modells) ist der minimale „Radius“ ε einer Umgebung um F , für den T zusammenbricht,

$$\varepsilon^*(T, F) = \inf_{\varepsilon} \langle b(\varepsilon; T, F) = \infty \rangle ,$$

(inf heisst infimum und kann mit Minimum übersetzt werden).

- f **Empirischer Bruchpunkt.** Wenn wir von den asymptotischen Begriffen wieder zu endlichen Stichproben zurückkehren, wird die vorhergehende mathematisch wirkende Funktion recht einfach und anschaulich. Wir gehen von einer Stichprobe x_1, x_2, \dots, x_n aus und fügen weitere q beliebige Werte $x_1^*, x_2^*, \dots, x_q^*$ als Beobachtungen hinzu. Wieder fragen wir, ob die Statistik T durch diese zusätzlichen Beobachtungen beliebig verfälscht werden kann, also wie gross $T(x_1, x_2, \dots, x_n, x_1^*, x_2^*, \dots, x_q^*) - T(x_1, x_2, \dots, x_n)$ werden kann. Wenn der Betrag dieser Differenz ∞ werden kann, ist der Zusammenbruch erfolgt, und der empirische Bruchpunkt ist der grösste Anteil $q/(n + q)$, für den dies nicht passiert.

Im Prinzip hängt der empirische Bruchpunkt von der Stichprobe ab. Häufig ist das aber dann doch nicht so. Beispielsweise ist anschaulich klar, dass das 10% gestutzte Mittel zusammenbricht, wenn der Anteil beliebig falscher Beobachtungen mehr als 10% beträgt, und zwar unabhängig von den gegebenen Beobachtungen x_1, x_2, \dots, x_n . Genauer: Für ein festes $n + q$ kann der Bruchpunkt nur ein Vielfaches von $1/(n + q)$ sein. Der empirische Bruchpunkt ist das kleinste Vielfache, das $\leq 10\%$ ist.

- g Die beiden Versionen des Bruchpunktes messen die Robustheit mit einer sehr groben Brille: Sie betrachten nur die „absolute Katastrophe“ eines totalen Zusammenbruchs. Für Ausreisser mit einem Anteil unterhalb des Bruchpunktes kann die Verfälschung recht gross werden; nur völlige Unbegrenztheit ist ausgeschlossen.

Der Vorteil dieser Masse besteht darin, dass sie recht einfach zu verstehen sind und keine Entscheidung darüber erfordern, was eine noch tolerierbare Verfälschung sein soll.

- h Wenn man sich dagegen absichern will, dass Ausreisser einen grossen Effekt auf ein Resultat haben, wird man eine sinnvolle Gross Error Sensitivity und einen genügend hohen Bruchpunkt verlangen. Die beiden Aspekte kann man sinnbildlich vergleichen mit der Sicherheit einer Brücke: Die Gross Error Sensitivity misst, wie stark die Brücke ins Vibrieren kommen kann, wenn ein relativ schwacher Wind bläst, und der Bruchpunkt misst, wie viel Wind es braucht, bis die Brücke einstürzen kann.

Beide Masse konzentrieren sich auf die schlimmstmögliche Windrichtung. Oft wird die Brücke auch bei stärkerem Wind standhalten, wenn er nicht gerade aus der Querrichtung bläst. Andererseits kann es schon praktisch unmöglich werden, über die Brücke zu fahren, wenn sie nicht am Einstürzen ist, aber gewaltig schwingt – der Bias $b\langle\varepsilon; T, F\rangle$ kann schon untolerierbar gross werden, wenn $\varepsilon < \varepsilon^*\langle T, F\rangle$ ist.

- i Kehren wir zu den kleinen Störungen ε zurück! Wir möchten dann eine möglichst kleine Gross Error Sensitivity. Man kann zeigen, dass für einen Lokationsparameter der Median die kleinst-mögliche Gross Error Sensitivity hat. Andererseits wissen wir, dass der Median eine wesentlich grössere Streuung hat als die optimale Schätzung, wenn die Beobachtungen normalverteilt sind.

Nochmals ein sinnbildlicher Vergleich: Um sich gegen das **Risiko** eines zu grossen Einflusses von Ausreissern abzusichern, muss man wie bei einer Versicherung eine **Prämie** zahlen – hier in Form eines Verlustes an Genauigkeit unter idealen Bedingungen.

- j Man kann nach den **guten Kompromissen** zwischen Absicherung und Prämie fragen. Prof. F. Hampel hat dieses Problem wie folgt formuliert: Gesucht ist die Schätzung, deren Gross Error Sensitivity höchstens gleich einer Schranke γ_0 ist und die unter dieser Nebenbedingung die kleinst-mögliche asymptotische Varianz aufweist. Er hat für dieses Problem auch allgemein die Lösung hergeleitet.

Für die Normalverteilung mit bekannter Varianz liefern die Huber-Schätzungen die Lösung des Problems. Figur 5.4.e zeigt, dass man für eine „Prämie“ von 5% mehr asymptotischer Varianz eine Schranke k von $k = 1.345$ erhält. Das entspricht einer Gross Error Sensitivity von 1.637. Der Median hat eine asymptotischer Varianz von 1.571 und eine Gross Error Sensitivity von 1.253. Der Kompromiss lohnt sich wohl!

5.7 Ausblick

- a Die Hauptzweck dieses Kapitels war die Einführung der Idee, eine asymptotische Normalverteilung als Näherung für die Verteilung einer Schätzung oder einer Teststatistik zu verwenden. Diese Idee wird in der Statistik immer wieder gebraucht. Da es dazu nützlich war, die Einflussfunktion einzuführen, haben wir sozusagen als Nebenprodukt Begriffe der robusten Statistik eingeführt. Dieses Thema wird uns in einem späteren Block beschäftigen, in dem es ausführlicher behandelt und vor allem auf die Regression angewandt wird.

-
- b **Mehrdimensionale Schätzungen.** Wir haben den Begriff der M-Schätzung zwar für mehrere Parameter eingeführt, in der nachfolgenden Theorie aber nur einen eindimensionalen Parameter betrachtet. Schon die Normalverteilung hat zwei Parameter. Um die Theorie für die Schätzung mehrerer Parameter zu formulieren, müssen wir zuerst die Begriffe der multivariaten Verteilungen, vor allem auch die multivariate Normalverteilung, kennen lernen, was in die Vorlesung multivariate Statistik gehört.
- c **Small Sample Asymptotics.** Es gibt Näherungen für Verteilungen von Statistiken T , die auf genaueren „Entwicklungen“ (analog Taylor-Entwicklung über das lineare Glied hinaus) beruhen. Sie produzieren als Näherungs-Verteilung nicht einfach Normalverteilungen und sind teilweise erstaunlich genau schon für sehr kleine Stichproben. (Stichworte: Sattelpunkt-Methoden, Edgeworth, Large Deviations, ...) Sie sind aber bisher nicht weit verbreitet.

Literaturverzeichnis

- Büning, H. und Trenkler, G. (1994). *Nichtparametrische statistische Methoden*, 2. Aufl., Walter de Gruyter, Berlin.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press. *includes 1 disk*
- Hartung, J. und Elpelt, B. (1997). *Multivariate Statistik: Lehr- und Handbuch der angewandten Statistik*, 6. Aufl., Oldenbourg, München.
- Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*, Wiley, N.Y.
- Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, Wiley Series in Probability and Statistics, 2nd edn, Wiley.
- Stahel, W. A. (2007). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 5. Aufl., Vieweg, Wiesbaden.