

Näherungen mit Bootstrap

Werner Stahel, Seminar für Statistik, ETH Zürich, 8. 4. 2009

Dieser Text kann dazu dienen, die Ideen des Bootstrap zusammenzufassen. Es fehlen hier Beispiele.

1 Fragestellung

- a **Schätzungen** sind Funktionen der Beobachtungen. Wenn ein parametrisches Modell eine Rolle spielt, sollen Schätzungen den **plausibelsten Wert** (unter irgendeinem Gesichtspunkt) für den Parameter angeben. Wenn kein parametrisches Modell ausdrücklich gegeben ist, dienen solche Funktionen der **Beschreibung der Beobachtungen**. Beispielsweise werden arithmetisches Mittel, empirische Standardabweichung, einfache Korrelation auch berechnet, ohne dass eine bestimmte Verteilungs-Familie für die Beobachtungen angenommen wird.
- b In wieder anderen Situationen wird zwar ein Modell postuliert, aber nicht alle Aspekte werden mit parametrischen Annahmen festgelegt. Ein Beispiel liefert eine lineare Regression $Y_i = \underline{x}_i^T \underline{\beta} + E_i$, bei der über die Fehler E_i nur angenommen werden soll, dass sie die gleiche Verteilung haben und unabhängig sind, aber keine Annahmen über diese Verteilung getroffen werden. Es interessiert eine Schätzung $\hat{\underline{\beta}}$ der Parameter $\underline{\beta}$. Solche Modelle, in denen ein Teil mittels Parameter formuliert wird, während ein anderer unspezifiziert bleibt, nennt man **semi-parametrische Modelle**. Wenn man für die Fehler $\mathcal{N}\langle 0, \sigma^2 \rangle$ annimmt, gilt σ^2 oft als lästiger Parameter (nuisance parameter). Man kann die ganze Fehler-Verteilung, wenn man sie nicht festlegen will, als „unendlich-dimensionalen lästigen Parameter“ auffassen.
- c In jedem Fall geben Funktionen der Beobachtungen das interessierende Resultat wieder. Auch Teststatistiken sind solche Funktionen. Wir wollen hier primär von Schätzungen reden.
- d „Eine Zahl ohne Genauigkeitsangabe ist wertlos.“
Die beste Art, eine solche **Genauigkeitsangabe** zu liefern, besteht in der Angabe eines Vertrauensintervalls. Dieses wiederum beruht auf einem Test, der durch eine Test-Statistik (meistens eine Schätzung) und den Annahmehereich gegeben ist. Der letztere wird durch die **Verteilung der Test-Statistik** unter einem genau spezifizierten Modell (der Nullhypothese) (und dem gewünschten Niveau α) bestimmt.
Was kann man retten, wenn man keine genauen Voraussetzungen über die Verteilung der Beobachtungen treffen will? – Erstaunlich viel! Davon handelt diese Lektion.
- e Bereits für den **t-Test** wurde die Nullhypothese nicht eindeutig spezifiziert: Der lästige Parameter σ^2 wurde offengelassen. Der Ausweg bestand darin, ihn zu schätzen. Man konnte dann die ursprüngliche Test-Statistik (\bar{X} respektive $\bar{Y}_2 - \bar{Y}_1$) standardisieren, und die Verteilung dieser neuen Grösse war vom lästigen Parameter unabhängig. – Die erste Idee, nämlich die Schätzung des lästigen Parameters, kann man auch auf den Fall übertragen, dass es sich um eine ganze „lästige Verteilung“ handelt.

2 Die Idee des Bootstrap

- a Die Verteilung \mathcal{L} einer Schätzung $T\langle X_1, X_2, \dots, X_n \rangle$ ist durch die Verteilung der Beobachtungen X_i bestimmt. Letztere können wir uns durch die kumulative Verteilungsfunktion G gegeben denken. Man kann deshalb vollständiger schreiben: $\mathcal{L} = \mathcal{L}\langle T; G \rangle$.

Für theoretische Untersuchungen der Eigenschaften von Schätzungen wird G als bekannt vorausgesetzt. Meistens wird angenommen, dass G aus einer „parametrischen Familie“ von Verteilungen stammt, $G = F_\theta$. T soll dann meistens eine Schätzung für θ sein.

In der Praxis ist G nicht bekannt. Um die Verteilung von T zu bestimmen, müssen wir ein geeignetes G in den Ausdruck $\mathcal{L}\langle T; G \rangle$ einsetzen. Dafür gibt es zwei prinzipiell verschiedene Möglichkeiten:

- Wir schätzen θ ja durch T , also liegt es nahe, die Verteilung aus dem parametrischen Modell, die dem geschätzten θ entspricht, einzusetzen,

$$\hat{\mathcal{L}} = \mathcal{L}\langle T; F_{\hat{\theta}} \rangle$$

- Wir ersetzen die unbekannt kumulative Verteilungsfunktion G durch die empirische kumulative Verteilungsfunktion \hat{G} ,

$$\hat{\mathcal{L}} = \mathcal{L}\langle T; \hat{G} \rangle$$

Diese einfache Idee heisst Bootstrap, genauer „**nichtparametrischer Bootstrap**“. Die erste Idee wird oft als „**Plug-in-Methode**“ bezeichnet: Wo unbekannt Parameter gebraucht werden, „stopft“ man eine Schätzung rein. Sie wird aber auch als „**parametrischer Bootstrap**“ bezeichnet.

Die Bestimmung von $\hat{\mathcal{L}} = \mathcal{L}\langle T; \hat{G} \rangle$ gelingt fast nie mit analytischen Methoden. Deshalb gehört zum nichtparametrischen Bootstrap fast immer auch die Simulation. Ein parametrischer Bootstrap ohne Simulation würde nicht als solcher wahrgenommen; da gehört es also immer dazu. Wir kommen darauf zurück.

- b Die Idee ist so einfach, dass vielen ein mulmiges Gefühl bleibt. Sie werden fragen: Weshalb haben wir denn einen solchen Aufwand mit parametrischen Modellen getrieben? Solche Fragen haben die mathematischen Statistiker seit dem Erscheinen des grundlegenden Artikels von Efron („Bootstrap Methods: Another Look at the Jackknife“, *Annals of Statistics* 7, 1-26, 1979) beschäftigt. Sie konnten die meisten mulmigen Gefühle klären.

Der Bootstrap ist eine sehr flexible Methode, um die Verteilung von Schätzungen zu schätzen. Das Wegfallen von Voraussetzungen kann zu einer kleineren Präzision führen als die Verfahren, die solche annehmen – wenn deren Voraussetzungen stimmen. Das teilt er mit anderen nicht-parametrischen und mit robusten Methoden.

3 Das Vorgehen

- a Es sei $n = 5$ und T das 20%-gestutzte Mittel

$$T\langle X_1, X_2, \dots, X_5 \rangle = \frac{1}{3}(X_{[2]} + X_{[3]} + X_{[4]}),$$

wobei die $X_{[k]}$ die geordneten Beobachtungen sind. Wenn die Verteilungsfunktion G der X_i gegeben ist, kann man prinzipiell die Verteilung $\mathcal{L}\langle T, G \rangle$ von T ausrechnen, sei es durch

- exakte Rechnung durch Bildung eines fünffachen Integrals, notfalls numerisch;
- Simulation;
- asymptotische Näherung.

Die erste Möglichkeit ist uns zu kompliziert, die dritte zu ungenau, es bleibt die Simulation.

- b Bei der **Simulation** geht man so vor:

Man erzeugt mittels Zufallszahlen mit Verteilung G eine Stichprobe vom gleichen Umfang n wie die beobachtete Stichprobe. Dann rechnet man das zugehörige T aus. Dies wiederholt man r Mal, wobei r meist in der Größenordnung von 500 bis 5000 liegt. Ein Histogramm der r erhaltenen Werte veranschaulicht dann die **simulierte Verteilung** von T .

Zur Erzeugung von „Pseudo“-Zufallszahlen gibt es in vielen Software-Paketen Funktionen, wenn die uniforme oder die Normalverteilung gewünscht werden. Will man eine andere Verteilung erhalten, dann ist dies mit Hilfe der gewünschten kumulativen Verteilungsfunktion F und uniform verteilten Zufallszahlen Z_i möglich: Man muss die letzteren nur mit der inversen Funktion F^{-1} transformieren, also $X_i^* = F^{-1}\langle Z_i \rangle$ bilden; diese folgen der gewünschten Verteilung. Figur 3.b zeigt die Idee. Wir schreiben X_i^* für simulierte Werte, um sie von den beobachteten Werten X_i unterscheiden zu können.

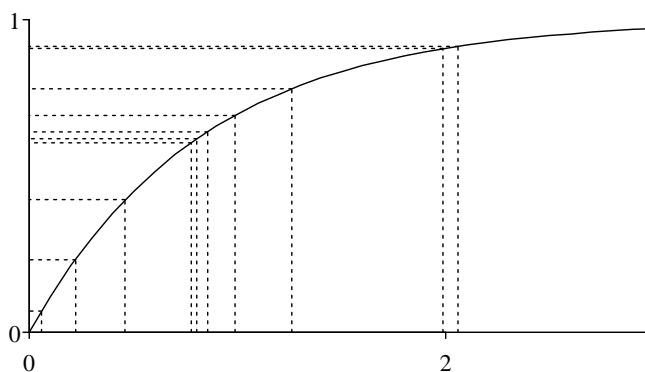


Abbildung 3.b: Erzeugung von Zufallszahlen mit gewünschter Verteilung

- c Die Simulation der Verteilung von T erfolgt nun für theoretische Untersuchungen mit bekannter Verteilung G und für den parametrischen Bootstrap mit $F_{\hat{\theta}}$.

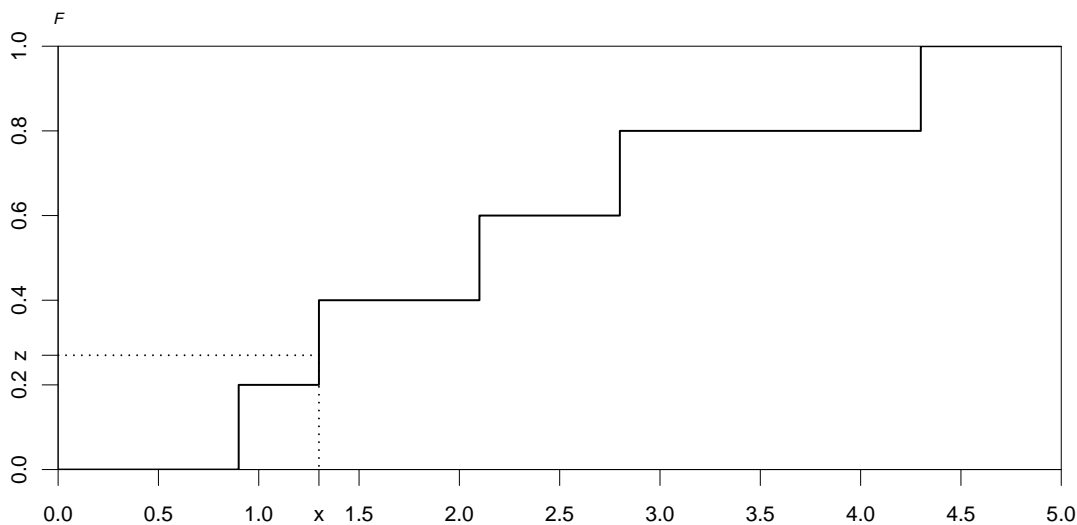


Abbildung 3.d: Simulation von Zufallszahlen gemäss \hat{G}

- d Für den nichtparametrischen Bootstrap, den wir hier ausführlicher betrachten wollen, ersetzen wir, wie gesagt, G durch \hat{G} . Die Beobachtungen seien 1.3, 2.8, 0.9, 4.3, 2.1. Die kumulative Verteilungsfunktion zeigt Figur 3.d.

Benützen wir also \hat{G} zur Simulation! Das führt zu einem merkwürdigen Vorgehen: Eine Zufallszahl entsprechend \hat{G} zu ziehen, entspricht der zufälligen Auswahl einer der Beobachtungen X_i , wie man an der Figur sehen kann. Eine ganze simulierte Stichprobe erhält man, indem man diesen Vorgang $n = 5$ mal wiederholt, das heisst, indem man 5 mal zufällig eine Beobachtung aus den 5 erhaltenen Werten zieht – mit Zurücklegen, denn eine Stichprobe zu simulieren bedeutet, n *unabhängige* Zufallszahlen zu ziehen. (Ein Ziehen ohne Zurücklegen würde immer zur ursprünglichen Stichprobe zurückführen.) In der simulierten Stichprobe treten also nur die Werte der beobachteten Stichprobe auf, die einen einmal, andere zweimal, ..., und einige gar nicht. Man nennt eine simulierte Stichprobe auch „**bootstrap sample**“.

- e Für jede simulierte Stichprobe wird T berechnet. Das ergibt eine simulierte Verteilung, die wir als geschätzte Verteilung von T benützen.

f **Das Vorgehen, allgemein formuliert.**

Gegeben ist eine Stichprobe x_1, x_2, \dots, x_n . Wir wollen die Verteilung von T schätzen.

Algorithmus.

Für $r = 1, \dots, \text{nrep}$:

Für $k = 1, \dots, n$: { Wähle zufällig ein $i \in \{1, \dots, n\}$ und behalte das entsprechende x_i (mit Wiederholungen), $:= \tilde{x}_k$ }

Berechne den zugehörigen Wert von $t_r = T(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$

Mit den Werten t_r , $r = 1, 2, \dots, \text{nrep}$, kann man ein Histogramm zeichnen, einen Median oder ein arithmetisches Mittel und eine Standardabweichung bestimmen als Simulationswerte für die entsprechenden Grössen der wahren Verteilung von T .

- g Wann funktioniert's? In den meisten Fällen, für die man sich in der Praxis interessiert. Aber nicht in allen.

Wann funktioniert's nicht? Beispielsweise, wenn T die kleinste Lücke zwischen zwei Beobachtungswerten ist. (Diese ist für alle möglichen bootstrap samples gleich 0 – bis auf

eines, die ursprüngliche Stichprobe nämlich.)

- h Es gibt natürlich Verfeinerungen dieses Rezepts und Anpassungen an bestimmte Modelle und Fragestellungen. Eine Erweiterung, die von Prof. Künsch an der ETH entwickelt wurde, ist ein Bootstrap für Zeitreihen, also für abhängige Beobachtungen.

4 Regression

- a Modell:

$$Y_i = \alpha + \beta x_i + E_i, \quad E_i \sim \mathcal{F}, \text{ unabhängig}$$

$T = \hat{\beta}$, eine Schätzung für die Steigung, kann Kleinste-Quadrate- oder robuste Schätzung sein. Ausgehend von den Beobachtungen $[x_1, y_1], [x_2, y_2], \dots, [x_n, y_n]$ wollen wir die Varianz von $\hat{\beta}$ bestimmen.

- b Es gibt zwei naheliegende Prozeduren für den Bootstrap:

Methode 1 (naive Methode).

Jedes Bootstrap sample besteht aus n zufällig ausgewählten Paaren $[x_{i_k}, y_{i_k}]$.

Methode 2 (Residuen bootstrappen).

Bestimme die Schätzwerte $\alpha^* := \hat{\alpha}$, $\beta^* = \hat{\beta}$ für die aktuellen Beobachtungen und bilde die Residuen $r_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$.

Bestimme jeweils ein bootstrap sample der Residuen und daraus y -Werte

$$\tilde{y}_k = \alpha^* + \beta^* x_k + r_{i_k} .$$

Bestimme $\hat{\beta}_\ell$ als Schätzwert für die simulierte „Stichprobe“ $[x_1, \tilde{y}_1], [x_2, \tilde{y}_2], \dots, [x_n, \tilde{y}_n]$.

Methode 2 entspricht dem Modell: Sie berücksichtigt, dass nicht $[x_i, y_i]$ zufällig ist, sondern E_i . Nur kennen wir E_i für die aktuellen Beobachtungen nicht. Deshalb ersetzen wir sie durch die Residuen. Eine weitere „Ungefähr“-Methode.

(... die man wenigstens noch verfeinern könnte, indem man berücksichtigt, dass die Varianzen von E_i und von R_i verschieden sind.)

Resultate: siehe freiwillige Übungen.

5 Was nützt der Bootstrap?

- a Nun haben wir also eine allgemeine Methode, um die **Verteilung einer Schätzung** zu schätzen. Wozu? Wir wollten die Verteilung unter einer allgemeinen Nullhypothese, um daraus einen **Annahmebereich** und schliesslich ein **Vertrauensintervall** zu erhalten.
- b Wo ist die Nullhypothese geblieben? – Sie legt ja jeweils die Verteilungsfunktion F_0 der Beobachtungen fest. Wenn wir F_0 durch die empirische Verteilungsfunktion \hat{G} ersetzen, haben wir sicher keine Chance, die Nullhypothese zu verwerfen; das hiesse ja, einen Widerspruch zwischen den Daten und — den Daten zu entdecken (genauer zwischen $T\langle X_1, X_2, \dots, X_n \rangle$ und $\hat{\mathcal{L}}\langle T; \hat{G} \rangle$).

- c Man behilft sich mit der üblichen „Plus-Minus-Regel“: Man „verwechselt“ das Vertrauensintervall mit einem „Streubereich“ – und das ist näherungsweise meistens in Ordnung! Man kann also
- mit dem Bootstrap die Standardabweichung der Verteilung von T (den „Standardfehler“) schätzen (\hat{s}_e) und $T \pm 2\hat{s}_e$ als Vertrauensintervall benutzen (gestützt auf die Annahme, dass T ungefähr normalverteilt sei);
 - von der geschätzten Verteilung das 2.5%- und das 97.5%-Quantil berechnen und diese als Grenzen des Vertrauensintervalls benutzen.
- d **Vertrauensintervall wofür eigentlich?** Zwei Antworten sind möglich:
- ... für den Erwartungswert von T ,
 - ... für den asymptotischen Wert von T .

Beides ist näherungsweise das Gleiche. Und die Vertrauensintervalle haben sowieso nur näherungsweise den richtigen Vertrauenskoeffizienten (die richtige Überdeckungs-Wahrscheinlichkeit).

- e **Der Bootstrap ermöglicht statistische „Inferenz“** (Vertrauensintervalle, Tests) **in sehr allgemeiner Weise, aber Vertrauenskoeffizienten oder Signifikanzniveaus werden immer nur näherungsweise eingehalten; üblicherweise ist die Näherung für grosse Stichproben genauer – und sogar besser als die Näherung mit der asymptotischen Normalverteilung!**
- L **Literatur:** In unserem Weiterbildungslehrgang wird das Buch von Davison and Hinkley (1997) empfohlen.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.