Fachbereich Mathematik & Informatik
Freie Universität Berlin
Carsten Hartmann, Ralf Banisch

6. Übung zur Vorlesung

Stochastik II

Sommersemester 2014

Aufgabe 1 (Zentraler Grenzwertsatz, 12 Punkte)
Sei (X_n) eine Folge von iid Zufallsvariablen mit Wertebereich $\{-1, 1\}$. Es gelte $\mathbb{P}(X_1 = 1) = \frac{1}{2} + \varepsilon$ und $\mathbb{P}(X_1 = -1) = \frac{1}{2} - \varepsilon$ für eine $0 < \varepsilon < \frac{1}{2}$. Setze $\mu := \mathbb{E}(X_1)$ und $\sigma^2 = \mathbb{V}(X_1)$ und betrachte

$$S_n = \sum_{k=1}^{n} \frac{X_k - \mathbb{E}(X_k)}{\sqrt{n}}.$$

a) Berechnen Sie μ und σ^2 sowie die charakteristische Funktion und die momenterzeugende Funktion von X_1 und S_n.

b) Formulieren und beweisen Sie den zentralen Grenzwertsatz für den vorliegenden Fall.

c) Entwerfen, implementieren und testen Sie ein Verfahren zur Berechnung von Realisierungen der Folge $(X_k)_{k=1,\dots,n}$. Verwenden Sie dieses Verfahren, um die Aussage des zentralen Grenzwertsatzes für $\varepsilon = 0.1$ und $n = 100, 1.000, 10.000$ zu testen. Verwenden Sie dabei jeweils ca. $m = 10.000$ Realisierungen und plotten Sie die empirische Verteilung. Überlegen Sie sich, wie Sie den Abstand zwischen der behaupteten Normalverteilung und der empirisch beobachteten Verteilung messen können.

d) Wiederholen Sie c) für $\varepsilon = 0.45$ und $\varepsilon = 0.49$. Was beobachten Sie? Wie können Sie sich Ihre Beobachtung erklären?

Hinweis: Wenn Sie m Realisierungen $X(\omega_1), ..., X(\omega_m)$ einer Zufallsvariable X vorliegen haben, dann können Sie mit Hilfe der Funktion `hist` in MATLAB das zugehörige Histogramm zeichnen, welches Ihnen für ausreichend große m eine Approximation der Verteilung von X liefert.

Lösung.

a) Wir haben $\mu = 2\varepsilon$ und $\sigma^2 = 1 - 4\varepsilon^2$. Die charakteristische Funktion von X_1 ist

$$\varphi_{X_1}(s) = \mathbb{E}[e^{isX_1}] = \left(\frac{1}{2} + \varepsilon\right) e^{is} + \left(\frac{1}{2} - \varepsilon\right) e^{-is}.$$

Die Momentenerzeugende Funktion von X_1 ist

$$M_{X_1}(s) = \mathbb{E}[e^{sX_1}] = \left(\frac{1}{2} + \varepsilon\right) e^{s} + \left(\frac{1}{2} - \varepsilon\right) e^{-s}$$

mit Definitionsbereich $D = \mathbb{R}$. Wir bestimmen die charakteristische Funktion von S_n und benutzen dabei, dass die X_i unabhängig und identisch verteilt sind:

$$\varphi_{S_n}(s) = \mathbb{E}\left[\prod_{i=1}^{n} e^{\frac{is}{\sqrt{n}}(X_i - \mu)}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{-\frac{is}{\sqrt{n}}\mu} e^{\frac{is}{\sqrt{n}}X_i}\right] = \left[\varphi_{X_1 - \mu}\left(\frac{s}{\sqrt{n}}\right)\right]^n$$

Damit ergibt sich schließlich

$$\varphi_{S_n}(s) = e^{-is\mu\sqrt{n}}\left[\left(\frac{1}{2}+\varepsilon\right)e^{\frac{is}{\sqrt{n}}} + \left(\frac{1}{2}-\varepsilon\right)e^{-\frac{is}{\sqrt{n}}}\right]^n.$$

Völlig analog folgt für die Momentenerzeugende Funktion von S_n:

$$M_{S_n}(s) = e^{-s\mu\sqrt{n}}\left[\left(\frac{1}{2}+\varepsilon\right)e^{\frac{s}{\sqrt{n}}} + \left(\frac{1}{2}-\varepsilon\right)e^{-\frac{s}{\sqrt{n}}}\right]^n.$$

b) Die Aussage des zentralen Grenzwertsatzes ist $S_n \xrightarrow{i.V.} \mathcal{N}(0,\sigma^2)$ für $n \to \infty$ (i.V. steht für 'in Verteilung'). Nach dem Stetigkeitssatz von Levy-Cramer ist das äquivalent zur punktweisen Konvergenz der charakteristischen Funktionen, also $\phi_{S_n}(s) \to \phi^*(s)$, wobei $\phi^*(s) = e^{-\frac{\sigma^2 s^2}{2}}$ die charakteristische Funktion der Normalverteilung darstellt. Wir zeigen die letzte Aussage. Zunächst haben wir, da $\mathbb{E}[|X_1|]$ und $\mathbb{E}[|X_1^2|]$ existieren, die Taylorentwicklung

$$\varphi_{X_1-\mu}(s) = 1 - \frac{\sigma^2 s^2}{2} + \mathcal{O}(s^3).$$

Also ist

$$\varphi_{S_n}(s) = \left[\varphi_{X_1-\mu}\left(\frac{s}{\sqrt{n}}\right)\right]^n = \left(1 - \frac{\sigma^2 s^2}{2n}\right)^n + R(n),$$

und wir bestimmen den dominanten Term des Restgliedes $R(n)$:

$$R(n) = n\mathcal{O}(n^{-3/2}) + \text{nachgeordnete Terme.}$$

Also $R(n) = \mathcal{O}(n^{-1/2})$, und $\varphi_{S_n}(s) \to e^{-\frac{\sigma^2 s^2}{2}}$ wie behauptet.

c) Man erzeugt sich $m \times n$ Realisierungen von X_1 z.B. so:

```
function [C]=iidRealize(eps,n,m)
X = zeros(n,m);
for i=1:n
    for j=1:m
        x=rand(1); %generates uniformly distributed pseudorandom number x\in[0,1]
        if x<(0.5+eps)
            X(i,j)=1;
        else
            X(i,j)=-1;
        end
    end
end
C=X;
end
```

Wir erzeugen uns daraus m Realisierungen der ZV S_n und erstellen das Histogram von S_n:

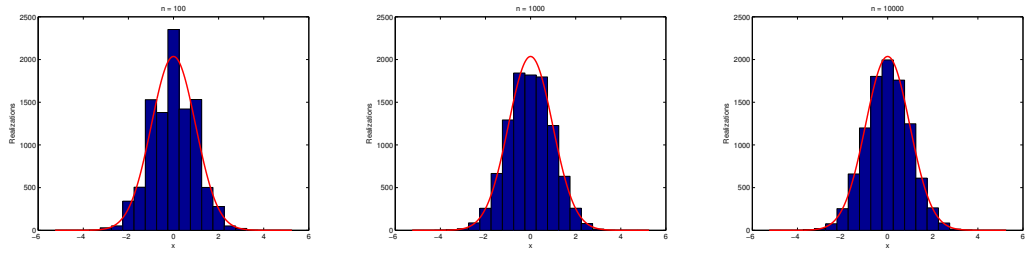
```
S = 1/sqrt(n)*(sum(X,1) - n*mean_X); %this should approximate the normal distribution
dx2 = 0.5; %histogram bin size
h = hist(S,-5:dx2:5);
```

Jetzt wollen wir dieses Histogram mit der erwarteten Normalverteilung vergleichen. Hier müssen wir allerdings auf die Normierungskonstante aufpassen, denn Histogramme sind i.A. nicht auf 1 normiert. Wir müssen die Verteilungen so normieren, dass 'Fläche unter dem Histogram' = 'Fläche unter der Normalverteilung' gilt:

```
    figure;
    hist(S,-5:dx2:5);
    Chist = sum(h*dx2); %the integral under the histogram-curve
    hold on
    hold all
    dx1 = 0.01;
    x = (-5-dx2/2):dx1:(5+dx2/2);
```

```
8       dist = 1/(sqrt(2*pi*Var_X))*exp(-x.^2/(2*Var_X)); %the normal distribution ...
            with mean mu and var sigma^2
9       Cdist = sum(dist*dx1); %the integral under the normal distribution (this ...
            should be 1)
10      p = plot(x,Chist/Cdist*dist,'r','linewidth',2);
11      xlabel('x');
12      ylabel('Realizations');
13      hold off;
```

Das sieht dann so aus:



Hier bekommt man schon einen ersten Eindruck von der Konvergenz gegen die erwartete Normalverteilung. Ein Problem beim Vergleichen von Histogrammen mit kontinuierlichen Verteilungen ist allerdings, dass Histogramme von der gewählten Auflösung, d.h. der Länge der bins $dx2$ abhängen - $dx2$ ist ein von uns frei wählbarer Parameter und darum ist jede Festlegung auf eine Binlänge zunächst einmal willkürlich.

Ein besseres Bild bekommen wir, wenn wir uns Verteilungsfunktionen anschauen. Die empirische Verteilungsfunktion von S_n ist

$$F_n(x) = \frac{1}{m}\sum_{j=1}^{m} \chi_{[1,x]}(S_n(j))$$

wobei $S_n(j)$ die j-te Realisierung der ZV S_n bezeichnet. F_n ist, im Gegensatz zum Histogramm, parameterfrei. Das vergleichen wir mit der Verteilungsfunktion der Normalverteilung, $F(x) = \mathbb{P}(\eta \leq x)$, $\eta \sim \mathcal{N}(0, \sigma^2)$:

```
1       F_n = zeros(1,length(x));
2       F = F_n;
3       for i=1:length(x)
4           F_n(i) = 1/m*sum(S<=x(i));
5           F(i) = sum(dx1*dist(1:i));
6       end;
7       figure;
8       plot(x,F);
9       hold on
10      plot(x,F_n,'r');
11      title('cumulative distribution function');
12      xlabel('x');
13      legend('F','F_n');
14      hold off
```

und erhalten das folgende Bild:

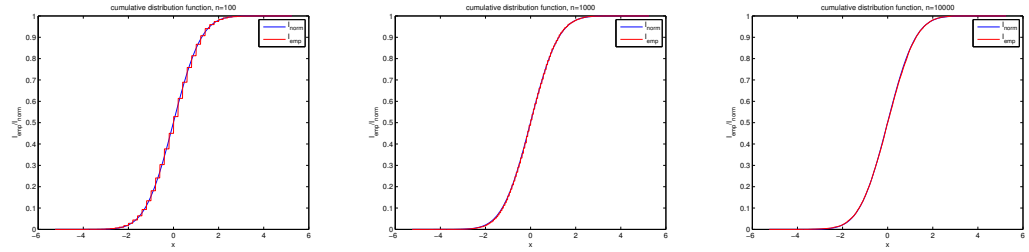


Figure 1: Empirische und tatsächliche Verteilungsfunktion, F_n und F, für $\varepsilon = 0.1$.

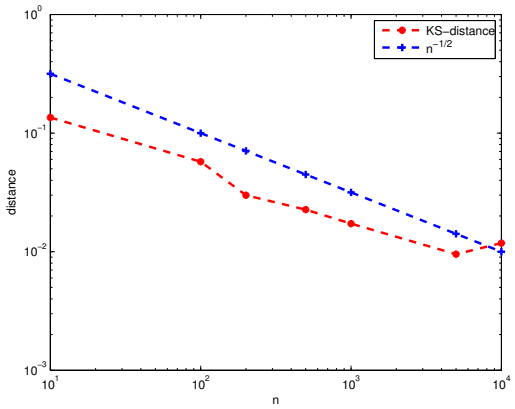
Die empirische Verteilungsfunktion gibt uns auch einen Abstandsbegriff zwischen empirischer und tatsächlicher Verteilung, den **Kolmogorov-Smirnov Abstand**

$$D_{KS} = \sup_x |F_n(x) - F(x)|.$$

In matlab ist das einfach `D_KS = max(abs(F_n-F))`. Wir betrachten uns D_{KS} als Funktion von n im loglog-Plot:

```
figure
loglog(N,D_KS,'--r*','linewidth',2);
hold on
loglog(N,1./sqrt(N),'--b+','linewidth',2);
xlabel('n');
ylabel('distance');
legend('KS-distance', 'N^{-1/2}');
hold off;
```

Das sieht dann so aus:



Der Kolmogorov-Smirnov Abstand D_{KS} fällt also ungefähr mit $n^{-1/2}$ ab.

d) Wir nehmen als Beispiel $\varepsilon = 0.49$. Die Varianz von X_1 ist jetzt viel kleiner, $\sigma^2 = 1 - 4\varepsilon^2 = 0.0396$. Unsere histogramme und Verteilungsfunktionen sehen jetzt folgendermaßen aus:

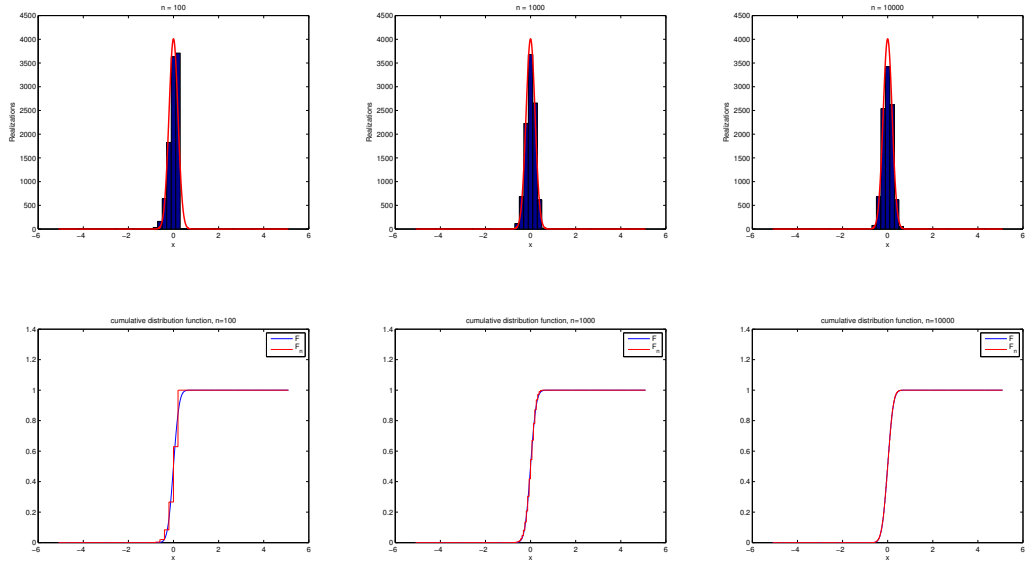


Figure 2: Empirische und tatsächliche Verteilungsfunktion, F_n und F für $\varepsilon = 0.49$.

Die Verteilungsfunktion ähnelt viel stärker einer Stufenfunktion als vorher. Zuletzt betrachten wir noch den KS-Abstand. Wie vorhin sehen wir einen Abfall mit $n^{-1/2}$, allerdings sind die Abstände etwa eine Größenordnung größer im Vergleich zu $\varepsilon = 0.1$. Das liegt daran, dass F für $\varepsilon = 0.49$ in einem kleinen Intervall sehr viel stärker variiert als für $\varepsilon = 0.1$, und solche Funktionen sind schwieriger zu approximieren.