Prof. Dr. Wolfgang Bischof
Prof. Dr. Friedrich Pukelsheim

Kurs 01363

Parametrische Statistik

LESEPROBE

Fakultät für
Mathematik und Informatik

FernUniversität in Hagen

7.8 Kolmogorov/Smirnov-Test

In Aufgabe 7.4.3 wurde eine Möglichkeit vorgestellt, wie man prüfen kann, ob Daten mit einer speziellen Normalverteilung modelliert werden können. Einen Test zum Prüfen der Hypothese, ob eine gewisse Verteilung vorliegt (gegen die Alternative, dass diese Verteilung nicht vorliegt) nennt man allgemein einen „Anpassungstest“.[19] Im vorliegenden Abschnitt erläutern wir einen berühmten Vertreter aus der Familie der Anpassungstests, den Kolmogorov/Smirnov-Test. Dabei kann ein signifikanter Nachweis - wie bei allen Verteilungstests - prinzipiell nur für die Hypothese geführt werden, dass die Daten *keine* Realisierungen dieser Verteilung sind. Ein Beweis der Hypothese, dass die Daten Realisierungen einer Verteilung sind, ist prinzipiell unmöglich, was allein schon deshalb einsichtig ist, weil es zu jeder Verteilungsfunktion F_0 eine andere Verteilungsfunktion F gibt, die sich beliebig wenig von F_0 unterscheidet.[20]

Wir nehmen nun an, dass n stochastisch unabhängige Beobachtungen $X_1, X_2, \ldots, X_n$ vorliegen, die alle die gleiche stetige Verteilungsfunktion F besitzen. Wie im Satz von Glivenko/Cantelli (1.19) definieren wir

$$D_n := \sup_{t\in\mathbb{R}} \left| \widehat{F}_{X_1,\ldots,X_n}(t) - F(t) \right|$$

als den Supremums-Abstand zwischen empirischer und theoretischer Verteilungsfunktion.

Wir zeigen nun, dass die Verteilung von D_n nicht von F abhängt.

Definiere dazu $U_i := F(X_i)$ für $i = 1, \ldots, n$. Dann sind die Zufallsvariablen $U_1, \ldots, U_n$ unabhängig und identisch $U_{(0;1)}$-verteilt[21] und es gilt:

$$\mathrm{P}\big(X_i \le t\big) = F(t) = \mathrm{P}\big(F(X_i) \le F(t)\big) = \mathrm{P}\big(F^{-1}(F(X_i)) \le t\big) = \mathrm{P}\big(F^{-1}(U_i) \le t\big),$$

wobei F^{-1} die Quantilfunktion[22] ist und das dritte Gleichheitszeichen wegen Satz 1.4.29 (a) gilt. Damit entspricht die Verteilung der Zufallsvariable

$$\widehat{F}_{X_1,\ldots,X_n}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i),$$

der Verteilung der Zufallsvariable

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}\left(F^{-1}\left(U_i\right)\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,F(t)]}\left(U_i\right) = \widehat{F}_{U_1,\ldots,U_n}\left(F(t)\right).$$[23]

[19] engl. goodness-of-fit test

[20] Man denke beispielsweise an die $N_{0;1}$- und t_{100}-Verteilung.

[21] wegen Satz 1.4.29 Teil (g) im dritten Anhang zu 1.4.

[22] Zur Definition der Quantilfunktion siehe den dritten Anhang zu 1.4.

[23] Beim ersten Gleichheitszeichen wurde wieder 1.4.29 (a) verwendet.

Weil F stetig ist und somit alle Werte im Intervall $(0;1)$ annimmt, ist der Abstand D_n genauso verteilt wie die Zufallsvariable

$$\sup_{t\in\mathbb{R}} \left| \widehat{F}_{U_1,\ldots,U_n}(F(t)) - F(t) \right| = \sup_{s\in(0;1)} \left| \widehat{F}_{U_1,\ldots,U_n}(s) - s \right|.$$

Der Abstand D_n zwischen der empirischen und theoretischen Verteilungsfunktion F hat also die gleiche Verteilung wie der Abstand D_n zwischen der empirischen und theoretischen Verteilungsfunktion bei n $U_{(0;1)}$-verteilten Zufallsvariablen, da für die Verteilungsfunktion der $U_{(0;1)}$-Verteilung $F_{U_{(0;1)}}(s) = s\mathbb{1}_{(0;1)}$ gilt. Damit haben wir gezeigt, dass die Verteilung von D_n nicht von der speziellen hypothetischen Verteilungsfunktion F abhängt, sondern mit den Zufallsvariablen $U_1, \ldots, U_n$ beschrieben werden kann.
Wir nehmen nun an, dass nicht nur endlich viele, sondern eine ganze Folge von u.i.v. Zufallsvariablen $X_1, X_2, \ldots$ gegeben ist. Die Reduktion auf die Rechteckvariablen $U_1, U_2, \ldots$ ist das zentrale Hilfsmittel, um die asymptotischen Verteilung beim Grenzübergang $n \to \infty$ zu bestimmen. Im Ergebnis erhält man für den mit $\sqrt{n}$ multiplizierten Abstand D_n im Limes die Kolmogorov-Verteilung K, die durch ihre Verteilungsfunktion $K(t)$ definiert ist.[24] Für alle $t > 0$ gilt:

$$\lim_{n\to\infty} \mathrm{P}(\sqrt{n}D_n \le t) = 1 - 2\sum_{k\ge 1}(-1)^{k-1}e^{-2k^2t^2} =: K(t).$$

Für das (einseitige) Testproblem

$$F = F_0 \quad \text{gegen} \quad F \neq F_0$$

lautet der Kolmogorov/Smirnov-Test:

$$\varphi(x_1,\ldots,x_n) = \begin{cases} 1 \\ 0 \end{cases} \text{ für } \sqrt{n}D_n(x_1,\ldots,x_n) \begin{cases} > \\ \le \end{cases} (1-\alpha)\text{-Quantil von } K.$$

Für das (zweiseitige) Testproblem:

$F \le F_0$ (d.h. F ist stochastisch größer[25] als F_0) gegen $F \ge F_0, F \neq F_0$.

Hierfür wird die Prüfgröße $\sqrt{n}D_n^+ = \sqrt{n}\sup_{t\in\mathbb{R}}\left(\widehat{F}_{X_1,\ldots,X_n}(t) - F(t)\right)$ benutzt (vgl. Abschnitt 1.20). [26] Die nun auftretende Limesverteilung heißt „Smirnov-Verteilung“; ihre Verteilungsfunktion ist gegeben durch $S(t) = 1 - e^{-2t^2}$ für

[24] vgl. Abschnitt 1.21

[25] siehe Abschnitt 1.20

[26] Für das einseitige Testproblem $F \ge F_0$ gegen $F \le F_0, F \neq F_0$ wird die Prüfgröße $\sqrt{n}D_n^- = \sqrt{n}\sup_{t\in\mathbb{R}} -\left(\widehat{F}_{X_1,\ldots,X_n}(t) - F(t)\right)$ benutzt.

$t > 0$.[27] Die ein- und zweiseitigen Fälle werden unter dem Namen Kolmogorov/Smirnov zusammengefasst.
In R kann man den ein- und zweiseitigen Kolmogorov/Smirnov-Test mit dem Kommando `ks.test` durchführen. Dabei kann über die Option `exact` gesteuert werden, ob zur Berechnung des p-Werts die exakte oder die asymptotische Verteilung von $\sqrt{n}D_n$ verwendet werden soll.

7.8.1 Aufgabe:
Beim Abpacken von Kartoffeln in Zehnkilosäcke kann das Normgewicht nicht genau eingehalten werden. Beim Wiegen von 20 Kartoffelsäcken ergaben sich folgende Werte (in kg):

9,92	*10,64*	*10,59*	*9,79*	*10,53*	*10,14*	*10,78*	*10,63*	*9,73*	*10,48*
10,76	*10,17*	*9,91*	*10,58*	*10,31*	*9,85*	*10,27*	*9,93*	*10,5*	*10,34*

Prüfen Sie mit dem Kolmogorov/Smirnov-Anpassungstest anhand obiger Beobachtungswerte, ob das Füllgewicht eines Kartoffelsacks durch eine auf dem Intervall (9,9;10,6)-gleichverteilte Zufallsvariable adäquat beschrieben werden kann. Verwenden Sie dabei sowohl die exakte als auch die asymptotische Verteilung von $\sqrt{n}D_n$.

7.8.2 Bemerkung: *Es ist zu beachten, dass der Test von Kolmogorov/Smirnov nur für den Test auf Gleichheit mit einer genau festgelegten Verteilungsfunktion F_0 funktioniert. Soll dagegen beispielsweise nur getestet werden, ob die Daten normalverteilt sind, dann lautet die Hypothese $H : F \in \{N_{\mu;\sigma^2} | \mu \in \mathbb{R}, \sigma^2 > 0\}$. Die Hypothese besteht also nicht aus einer (einzigen) Verteilung F_0, sondern aus der Familie aller Normalverteilungen. Es ist verführerisch, für Erwartungswert und Varianz ihre Schätzwerte einzusetzen, um mit $\widehat{F}_0 = N_{\overline{x},s^2}$ eine einfache Hypothese zu „erzwingen“ und einer (naiven) Anwendung des Kolmogorov/Smirnov-Tests den Weg zu ebnen. Dies führt in der Regel zu einem zu großen p-Wert. Speziell für den Test auf Normalverteilung kann stattdessen der Lilliefors-Test, der eine Variante des Kolmogorov/Smirnov-Tests darstellt, verwendet werden. Er benutzt die gleiche Prüfgröße wie der Kolmogorov/Smirnov-Test, berücksichtigt aber, dass die Parameter aus der Stichprobe geschätzt wurden.*

7.8.3 Beispiel: *Testen Sie, ob die Daten aus Aufgabe 7.8.1 normalverteilt sind.*

[27] vgl. wieder Abschnitt 1.21

Da die Parameter der Normalverteilung nicht bekannt sind, müssen sie aus der Stichprobe geschätzt werden. Folglich ist die Lilliefors-Variante des Kolmogorov/Smirnov-Tests zu verwenden. Diese ist im Paket `nortest` *enthalten.*

```
library("nortest")
lillie.test(gew)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  gew
## D = 0.15837, p-value = 0.2095

# Achtung: Der KS-Test mit geschätzen Parametern führt
# in die Irre
ks.test(gew, "pnorm",mean(gew),sd(gew))

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  gew
## D = 0.15837, p-value = 0.6409
## alternative hypothesis: two-sided
```

Man sieht, dass der Kolmogorov/Smirnov-Test zwar die gleiche Prüfgröße $D = \sqrt{n}D_n$ *verwendet, aber einen viel zu großen p-Wert liefert.*

7.9 Cramér/von Mises-Test

Beim Kolmogorov/Smirnov-Test wurde der maximale Abstand zwischen Verteilungsfunktion und empirischer Verteilungsfunktion herangezogen, um die empirische Verteilungsfunktion mit einer vorgegebenen Verteilungsfunktion zu vergleichen. Im Gegensatz dazu wird beim Cramér/von Mises-Test der L^2-Abstand zwischen der empirischen Verteilungsfunktion und der zu testenden Verteilungsfunktion F_0 herangezogen. Dabei beschränken wir uns auf stetige Verteilungsfunktionen F_0, die fast sicher differenzierbar sind.
Wenn die Beobachtungen $X_1, \ldots, X_n$ gemäß der Verteilungsfunktion F_0 verteilt sind, hängt auch die Verteilung des L^2-Abstands nicht von F_0 ab, wie

folgende Gleichheitskette[28] zeigt. Dabei wird die Notation und die bereits bewiesene Gleichheitskette aus dem vorherigen Abschnitt verwendet.

$$\int\limits_{\mathbb{R}} \left(\widehat{F}_{X_1,\ldots,X_n}(t) - F_0(t)\right)^2 f_0(t)\,dt =$$

$$= \int\limits_{\mathbb{R}} \left(\widehat{F}_{U_1,\ldots,U_n}(F_0(t)) - F_0(t)\right)^2 f_0(t)\,dt = \int\limits_0^1 \left(\widehat{F}_{U_1,\ldots,U_n}(s) - s\right)^2 ds,$$

wobei X_1 die Verteilungsfunktion F_0 besitze und $f_0(t) := \frac{d}{dt}F_0(t)$ die Ableitung von F_0 und damit die Lebesgue-Dichte von P^{X_1} ist. Das letzte Gleichheitszeichen gilt wegen der Substitution $s := F_0(t)$ und der Substitutionsregel.
Analog zum Satz 1.21 von Kolmogorov/Smirnov kann auch die asymptotische Verteilung des L^2-Abstands angegeben werden.

7.9.1 Satz *Seien $Z_1, Z_2, \ldots$ eine Folge unabhängig und identisch standardnormalverteilter Zufallsvariablen. Dann konvergiert die Verteilung des n-fachen des L^2-Abstands zwischen empirischer und theoretischer Verteilungsfunktion*

$$T(X_1,\ldots,X_n) := n \int\limits_{\mathbb{R}} \left(\widehat{F}_{X_1,\ldots,X_n}(t) - F_0(t)\right)^2 f_0(t)\,dt$$

für $n \to \infty$. Die Grenzverteilung geht auf Richard von Mises[29] zurück und entspricht der Verteilung der (fast-sicher konvergierende) Reihe $\sum\limits_{k=1}^{\infty} \frac{Z_k^2}{k^2\pi^2}$. Die Grenzverteilung hängt also insbesondere nicht von F_0 ab.

Damit kann der Cramér/von Mises-Test konstruiert werden. Seien dazu X_1, $\ldots, X_n$ u.i.v. gemäß einer stetigen Verteilungsfunktion F. Für das Testproblem

$$F = F_0 \quad \text{gegen} \quad F \neq F_0$$

lautet der Cramér/von Mises-Test:

$$\varphi(x_1,\ldots,x_n) = \begin{cases} 1 & \\ 0 & \end{cases} \text{für } T(x_1,\ldots,x_n) \begin{cases} > \\ \leq \end{cases} (1-\alpha)\text{-Quantil der Grenzvtlg.}$$

[28] Alle Gleichungen sind als Gleichheiten der Verteilung der linken und rechten Seite zu verstehen.

[29] Richard von Mises, Wahrscheinlichkeitsrechnung, Teubner, Leipzig, 1931. Siehe auch Ralph B. D'Agostino, Michael A. Stephens (Editors) Goodness-of-Fit Techniques, Marcel Dekker, Inc., New York and Basel, 1986, S. 100ff.

In R ist der Cramér/von Mises-Test beispielsweise über das Kommando `cvm.test` im Goodness-of-Fit-Paket `goftest` aufrufbar.

7.9.2 Beispiel: *Testen Sie in der Situation von Aufgabe 7.8.1 mit dem Cramér/von Mises-Test, ob die Annahme, dass das Füllgewicht eines Kartoffelsacks durch eine auf dem Intervall (9,9;10,6)-gleichverteilte Zufallsvariable adäquat beschrieben werden kann, mit den obigen Beobachtungswerten verträglich ist.*

Nach dem Laden des Pakets `goftest` *führt der Befehl*

```
cvm.test(gew, "punif", 9.9, 10.6)
```

zur Ausgabe

```
Cramer-von Mises test of goodness-of-fit
Null hypothesis: uniform distribution

data:  gew
omega2 = 0.50493, p-value = 0.03763
```

Anders als beim Kolmogorov/Smirnov-Test kann die Nullhypothese nun also auf einem Signifikanzniveau von 5% abgelehnt werden. Die Daten sind auffallend unverträglich mit der $U_{(9,9;10,6)}$-Verteilung. Es sollte ein anderes Modell gewählt werden.

7.10 Normal-Scores-Test zum Prüfen einer Normalverteilungsannahme

In der Praxis möchte man häufig überprüfen, ob eine Normalverteilungsannahme für die beobachtete Verteilung von numerischen Daten gerechtfertigt ist oder ob die Verteilung der Daten signifikant von der Normalverteilung abweicht. Für diesen Spezialfall stellen wir nun einen weiteren Verteilungstest, den Normal-Scores-Test, vor. Er basiert auf der Korrelation zwischen den sortierten beobachteten Datenwerten und den Werten, die man dafür unter der Normalverteilungsannahme erwarten würde.
Als Modell betrachten wir wieder u.i.v. reellwertige Zufallsvariablen $X_1, \ldots, X_n$ mit stetiger Verteilungsfunktion F. Wenn man mit $\mathcal{N} := \{N_{\mu;\sigma^2} : \mu \in \mathbb{R}, \sigma^2 > 0\}$ die Menge der Normalverteilungen bezeichnet, lautet das Testproblem: $H_0 : F \in \mathcal{N}$ gegen $F \notin \mathcal{N}$.

Sei ohne Einschränkung $X_i \neq X_j$ für $i \neq j$. Definiere

$$\begin{aligned} Y_1 &:= \min\{X_1, \ldots, X_n\} \\ Y_2 &:= \min(\{X_1, \ldots, X_n\} \setminus \{Y_1\}) \\ &\ldots \\ Y_n &:= \min(\{X_1, \ldots, X_n\} \setminus \{Y_1, \ldots, Y_{n-1}\}) \end{aligned}$$

Der Vektor $X_\uparrow := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ wird als „Ordnungsstatistik“ bezeichnet.

Was erwarten wir für $X_\uparrow$ unter der Nullhypothese?

Zur Klärung dieser Frage betrachten wir den Spezialfall der Standardnormalverteilung. Seien dazu $Z_1, \ldots, Z_n$ stochastisch unabhängige und identisch standardnormalverteilte Zufallsvariablen mit der zugehörigen Ordnungsstatistik $Z_\uparrow$. D.h. $Z_{\uparrow i}$ ist die i-kleinste Beobachtung unter $\{Z_1, \ldots, Z_n\}$. Es erweist sich als zweckmäßig, statt dem Erwartungswert von $Z_\uparrow$ den Median zu untersuchen.[30] Setze $a_i := \operatorname{med}[Z_{\uparrow i}]$, das ist der Median der i-kleinsten der n Zufallsvariablen $Z_1, \ldots, Z_n$. Auch wenn die Zufallsvariablen $Z_1, \ldots, Z_n$ natürlich den Median Null haben, ist a_i im Allgemeinen verschieden von Null. Denn wenn wir die Zufallsvariablen $Z_1, \ldots, Z_n$ sortieren, liegt der Median von $Z_{\uparrow 1}$ im Negativen und der Median von $Z_{\uparrow n}$ im Positiven.

Wegen den Eigenschaften der Normalverteilung erwarten wir, dass die Punkte $(x_{\uparrow i}, a_i)$ unter der Nullhypothese auf einer Geraden liegen, also linear korreliert sind. Das motiviert den Normal-Scores-Test:

$$\varphi(X_1, \ldots, X_n) \;=\; \begin{cases} 1 & \text{für } \operatorname{Corr}[X_\uparrow, a] \\ 0 & \end{cases}$$

$$\begin{cases} > \\ \leq \end{cases} (1-\alpha)\text{-Quantil der Vtlg. von } \operatorname{Corr}[X_\uparrow, a] \text{ unter } H_0.$$

Dabei ist $\operatorname{Corr}[X_\uparrow, a]$ die Stichprobenkorrelation zwischen $X_\uparrow$ und a.

Zur näherungsweisen Berechnung der Medianwerte $a_1, \ldots, a_n$ wurden verschiedene Methoden vorgeschlagen. Zwei davon werden im nächsten Abschnitt erläutert.

Der Normal-Scores-Test kann in R mit dem Kommando `ppccTest`[31] aus dem Paket `ppcc` durchgeführt werden. Dabei kann über die Option `ppos` angegeben werden, mit welcher Methode die Mediane approximiert werden. Beispiele dazu werden am Ende des nächsten Abschnitts vorgestellt.

[30] Auf den Erwartungswert der Ordungsstatistik wird in Abschnitt 7.11 eingegangen.

[31] „ppcc“ steht für *propability plot correlation coefficient*.

7.11 Approximation der mittleren Ordnungsstatistiken unter Normalverteilungsannahme

Der folgende Satz dient dazu, die für den Normal-Scores-Test benötigten Mediane a_i näherungsweise zu berechnen.

7.11.1 Satz *Seien $Z_1, \ldots, Z_n$ stochastisch unabhängige und identisch standardnormalverteilte Zufallsvariablen mit Ordnungsstatistik $Z_\uparrow$. Außerdem seien die Zufallsvariablen $U_1, \ldots, U_n$ stochastisch unabhängig und identisch auf $(0;1)$ gleichverteilt mit Ordnungsstatistik $U_\uparrow$. Schließlich bezeichne c_i den Median der Verteilung von $Z_{\uparrow i}$ und d_i den Median der Verteilung von $U_{\uparrow i}$. Dann gelten:*

(a) Für alle $i = 1, \ldots, n$ ist $c_i = \Phi^{-1}(d_i)$.

(b) Für alle $i \geq \frac{n+1}{2}$ ist

$$\frac{i}{n+1} \leq d_i \approx \frac{i - \frac{3}{8}}{n + \frac{1}{4}} \leq \frac{i-1}{n-1}.$$

BEWEIS:

(a) Reduziere zunächst das Problem von der Normalverteilung auf eine kontinuierliche Gleichverteilung: $U_i := \Phi(Z_i)$ ist nach Satz 1.4.29 (g) $U_{(0;1)}$-verteilt. Für $z \in \mathbb{R}^n$ gilt wegen der Isotonie von Φ

$$\Phi(z_{\uparrow i}) = \Phi\left((z_1, \ldots, z_n)_{\uparrow i}\right) = (\Phi(z_1) \ldots \Phi(z_n))_{\uparrow i}.$$

Somit folgt

$$Z_{\uparrow i} \leq \Phi^{-1}(d_i) \;\Leftrightarrow\; \Phi(Z_{\uparrow i}) \leq d_i \;\Leftrightarrow\; U_{\uparrow i} \leq d_i.$$

Damit folgt, dass $\Phi^{-1}(d_i)$ der Median von $\mathrm{P}^{Z_{\uparrow i}}$ ist, denn

$$\begin{aligned} \mathrm{P}\left(Z_{\uparrow i} \leq \Phi^{-1}(d_i)\right) &= \mathrm{P}(U_{\uparrow i} \leq d_i) = 0.5, \\ \mathrm{P}\left(Z_{\uparrow i} \geq \Phi^{-1}(d_i)\right) &= \mathrm{P}(U_{\uparrow i} \geq d_i) = 0.5. \end{aligned}$$

(b) Die zweite Aussage werden wir in insgesamt sechs Einzelschritten nachweisen.

(I) Die Ordnungsstatistik $U_{\uparrow i}$ ist $Beta_{i;n-i+1}$-verteilt[32], d.h. für $t \in (0;1)$ gilt:

$$\mathrm{P}\left(U_{\uparrow i} \leq t\right) = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)} \int_0^t u^{i-1}(1-u)^{n-i}\, du.$$

(II) Nach Abschnitt 1.4 gilt $\mathrm{E}\left[U_{\uparrow i}\right] = \frac{i}{n+1}$.

Setzt man die Ableitung der Dichte der $Beta_{i;n-i+1}$-Verteilung

$$\begin{aligned} \frac{d}{du} u^{i-1}(1-u)^{n-i} &= u^{i-2}(1-u)^{n-i-1}\big((i-1)(1-u) - (n-i)u\big) \\ &= u^{i-2}(1-u)^{n-i-1}\big((i-1) - (n-1)u\big), \end{aligned}$$

gleich Null, erhält man $\frac{i-1}{n-1}$ als Gipfelwert von $U_{\uparrow i}$.

(III) Bei Anwendung der Mean-Median-Mode-Ungleichung[33] auf die $Beta_{i;n-i+1}$-Verteilung mit $1 < n-i+1 < i \Leftrightarrow i \geq \frac{n+1}{2}$, erhält man für alle $i \geq \frac{n+1}{2}$:

$$\mathrm{E}\left[U_{\uparrow i}\right] = \frac{i}{n+1} \leq \mathrm{med}\left(U_{\uparrow i}\right) \equiv d_i \leq \mathrm{mode}\left(U_{\uparrow i}\right) = \frac{i-1}{n-1}$$

(IV) Mit der Hilfsfunktion $g(\lambda) := \frac{i-\lambda}{n-2\lambda+1}$ lässt sich die letzte Ungleichungskette so ausdrücken

$$g(0) = \frac{i}{n+1} \leq \frac{i-1}{n-1} = g(1).$$

Da $g(\lambda)$ wegen $g'(\lambda) = \frac{2}{(n-2\lambda+1)^2}\left(i - \frac{n+1}{2}\right) \geq 0$ isoton ist, gilt

$$g(0) = \frac{i}{n+1} \leq g(\lambda) \leq \frac{i-1}{n-1} = g(1) \;\Leftrightarrow\; \lambda \in [0;1].$$

(V) Nach dem Zwischenwertsatz wissen wir, dass für jedes n und für jedes $i \geq \frac{n+1}{2}$ ein $\lambda \in [0;1]$ existiert, so dass $g(\lambda) = \frac{i-\lambda}{n-2\lambda+1} = d_i$. Dieses λ hängt jedoch von i und n ab.

Auf der Suche nach einem globalen λ, das für alle i und n anwendbar ist, wurden verschiedene λ-Werte vorgeschlagen (siehe Bemerkung nach dem Beweis). Üblicherweise wird heute $\lambda = \frac{3}{8}$ gewählt.

---

[32] Siehe Georgii, Stochastik, de Gruyter, 2015, 5. Auflage, S.47 ff.

[33] Siehe J. Th. Runnenburg (1978), Mean, median, mode, Statistica Neerlandica, 32 (2): 73–79 und van Zwet, WR (1979), Mean, median, mode II, Statistica Neerlandica, 33 (1): 1–5.

(VI) Wegen $\mathrm{P}(U_{\uparrow n} \leq t) = t^n$ für $t \in [0;1]$, gilt ferner $d_n = \left(\frac{1}{2}\right)^{\frac{1}{n}}$. Damit kann man die Ergebnisse aus Symmetriegründen wie folgt zusammenfassen:

$$d_i \begin{cases} = \left(\frac{1}{2}\right)^{\frac{1}{n}} & \text{für } i = n \\ \approx \frac{i-\frac{3}{8}}{n+\frac{1}{4}} = 1 - d_{n-i+1} & \text{für } i = 2, \ldots, n-1 \\ = 1 - \left(\frac{1}{2}\right)^{\frac{1}{n}} & \text{für } i = 1 \end{cases}$$

Dies beinhaltet die zweiten Teilaussage. #

7.11.2 Bemerkung: *Historisch wurden einige Varianten für die Wahl von λ vorgeschlagen. Hier sollen nur zwei erwähnt werden: Filliben*[34] *verwendete $\lambda = 0,3175$ und für d_1 und d_n die in der Zusammenfassung genannten Werte. Blom*[35] *verwendete das im Satz genannte $\lambda = \frac{3}{8}$ (auch für d_1 und d_n).*
Mit dem Befehl `ppPositions` *aus dem* `ppcc`*-Paket erhält man die d_i-Werte für die oben genannten Varianten über das Setzen der Option* `method` *auf* `Filliben` *bzw.* `Blom`*.*

7.11.3 Bemerkung: *Im Normal-Scores-Test könnten statt der Mediane der Ordungsstatistik auch die erwarteten Ordnungsstatistiken verwendet werden. Diese erhält man in R mit dem Befehl* `normOrder` *aus dem Paket* `SuppDists` *oder alternativ mit dem Befehl* `evNormOrdStats` *aus dem sehr umfangreichen Paket* `EnvStats`*, das einige verschiedene Approximationsoptionen bietet. Tatsächlich liegen die erwarteten Ordnungsstatistiken so nahe an den mit der Blom-Methode approximierten Medianen (siehe auch folgende Übersichtstabelle), dass es sich nicht lohnt, einen Normal-Scores-Test, der die erwarteten Ordnungsstatistiken verwendet, zu implementieren.*

In folgender Übersichtstabelle sind die *wahren*[36] und die mit den Methoden von Filliben und Blom geschätzten Mediane sowie die Erwartungswerte der Ordnungsstatistiken der Standardnormalverteilung für zehn Beobachtungen aufgelistet. Die entsprechenden Werte für die Ordnungsstatistiken $Z_{\uparrow 6}, \ldots, Z_{\uparrow 10}$ ergeben sich aus Symmetriegründen gerade als das Negative der aufgelisteten Werte also $\mathrm{med}\,(Z_{\uparrow 10}) = -Z_{\uparrow 1} = 1,499$ usw.

[34] Filliben, J. J. (February 1975), The Probability Plot Correlation Coefficient Test for Normality, Technometrics, American Society for Quality, 17 (1): 111–117.

[35] Blom, G. (1958), Statistical estimates and transformed beta variables, New York: John Wiley and Sons.

[36] Die *wahren* Mediane wurden durch Simulation von 100 Millionen standardnormalverteilten Zufallszahlen ermittelt.

	Median	Filliben	Blom	Erwartungswert
$Z_{\uparrow 1}$	-1.499	-1.499	-1.547	-1.539
$Z_{\uparrow 2}$	-0.985	-0.985	-1.000	-1.001
$Z_{\uparrow 3}$	-0.648	-0.647	-0.655	-0.656
$Z_{\uparrow 4}$	-0.372	-0.371	-0.375	-0.376
$Z_{\uparrow 5}$	-0.121	-0.121	-0.123	-0.123

Man sieht, dass Fillibens Näherung für den Median sehr gut ist. Bloms Werte liegen näher am Erwartungswert als am Median.

7.11.4 Beispiel: *Führen Sie für die Daten aus Aufgabe 7.8.1 Normal-Scores-Tests durch. Verwenden Sie als Näherung für die Mediane der Teststatistik die Varianten von Filliben und Blom. Vergleichen Sie die p-Werte untereinander und mit dem p-Wert aus Beispiel 7.8.3.*

```
## Zwei Varianten des Normal-Scores-Test
##
library("ppcc")
gew <- c(9.92, 10.64, 10.59,  9.79, 10.53, 10.14, 10.78, 10.63,
         9.73, 10.48, 10.76, 10.17,  9.91, 10.58, 10.31,  9.85,
         10.27, 9.93, 10.5, 10.34)
# 1. Variante: Normal-Scores-Test nach Filliben
ppccTest(gew, ppos="Filliben")

##
##  Probability Plot Correlation Coefficient Test
##
## data:  gew
## ppcc = 0.97268, n = 20, p-value = 0.2667
## alternative hypothesis: gew differs from a Normal distribution

# Anmerkung: Teststatistik entspricht Stichprobenkorrelation
# zwischen den Gewichtsdaten und den nach Filliben geschätzten
# Medianen der Standardnormalverteilung
cor(sort(gew), qnorm(ppPositions(20, method="Filliben")))

## [1] 0.9726806

#
# 2. Variante: Normal-Scores-Test mit nach Blom
ppccTest(gew, ppos="Blom")
```

```
##
##  Probability Plot Correlation Coefficient Test
##
## data:  gew
## ppcc = 0.97149, n = 20, p-value = 0.2422
## alternative hypothesis: gew differs from a Normal distribution

# diesmal entspricht die Teststat. der Stichprobenkorrelation
# zwischen den Gewichtsdaten und den nach Blom geschätzten
# Medianen der Standradnormalverteilung
cor(sort(gew), qnorm(ppPositions(20, method="Blom")))

## [1] 0.9714904

# Die p-Werte sind ähnlich groß und liegen knappe sechs bzw.
# drei Prozentpunkte über dem p-Wert, der in Beispiel 7.8.3
# mit dem Lilliefors-Tests berechnet wurde.
```

7.12 QQ-Plots

Ein Quantile-Quantile-Plot, kurz QQ-Plot, ist ein exploratives, grafisches Werkzeug, um zu prüfen, ob die Verteilung von Beobachtungswerten mit einer gegebenen Verteilung verträglich ist.
Dazu ordnet man die Beobachtungswerte $\{b_i\}_{i=1,\ldots,n}$ der Größe nach und vergleicht sie mit dem Erwartungswert oder Median a_i der i.-größten von n Beobachtungen der theoretischen Verteilung, in dem man die Paare (a_i, b_i) in ein Koordinatensystem aufträgt und prüft, ob die Punkte ungefähr auf einer Geraden liegen.
Praktisch führt man QQ-Plots in R mit dem Befehl `qqnorm` durch, um zu prüfen ob die Daten mit einer Normalverteilung verträglich sind. Dabei wird bis zu einem Stichprobenumfang von $n = 10$ die im letzten Abschnitt besprochene Näherung der a_i nach Blom verwendet, also $a_i = \Phi^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right)$. Für $n > 10$ wird die Näherung nach Hazen, das ist $\lambda = 0,5$, also $a_i = \Phi^{-1}\left(\frac{i-\frac{1}{2}}{n}\right)$ verwendet. Führt man nach dem `qqnorm`-Befehl den Befehl `qqline` aus, so wird eine Hilfslinie gezeichnet, die es erleichtert zu prüfen, ob alle Punkte auf einer Geraden liegen (Details dazu findet man in der Hilfefunktion von `qqline`.).