

1.8 Kolmogorov-Smirnov-Test auf Normalverteilung

Der Kolmogorov-Smirnov-Test ist einer der klassischen Tests zum Überprüfen von Verteilungsvoraussetzungen. Der Test vergleicht die Abweichungen der empirischen Verteilungsfunktion mit der theoretischen Verteilungsfunktion, d.h. in unserem Fall der Normalverteilungsfunktion. Wir gehen wieder davon aus, dass die Werte der Stichprobe x_1, x_2, \dots, x_n Realisierungen von unabhängig und identisch verteilter Zufallsvariablen X_1, X_2, \dots, X_n sind.

Getestet wird:

H_0 : Die Daten stammen aus einer normalverteilten Grundgesamtheit gegen

H_1 : Die Daten stammen aus keiner normalverteilten Grundgesamtheit

bzw.

H_0 : Die Zufallsvariablen X_i haben die Verteilungsfunktion F_0 gegen

H_1 : Die Zufallsvariablen X_i haben nicht die Verteilungsfunktion F_0

F_0 ist in unserem Fall die Verteilungsfunktion der $N(\bar{x}, s^2)$ -Verteilung.

Es folgt ein Beispiel:

v1
167
163
155
167
161
177
173
179

Über → **Univariate Statistik** und → **Kolmogorov-Smirnov-Test** erhalten Sie den folgenden Output:

Kolmogorov-Smirnov-Test (auf Normalverteilung)

Beobachtung	Funktionswert empirische Verteilungsfunktion	Funktionswert Verteilungsfunktion Normalverteilung ⁽¹⁾	Betrag Differenz 1	Betrag Differenz 2
155	0.125	0.060149	0.064851	0.060149
161	0.25	0.205409	0.044591	0.080409
163	0.375	0.281374	0.093626	0.031374
167	0.625	0.463594	0.161406	0.088594
173	0.75	0.738812	0.011188	0.113812
177	0.875	0.870143	0.004857	0.120143
179	1	0.914775	0.085225	0.039775

⁽¹⁾ N(167.75, 67.357143)-Verteilung

Maximale Abweichung: 0.161406
 Prüfgröße zum K-S-Test: 0.456525
 asymptotischer p-Wert (zweiseitig): 0.98525

Die empirische Verteilungsfunktion der Stichprobe x_1, x_2, \dots, x_n ist definiert über die kumulierten relativen Häufigkeiten:

$$F_{\text{emp}}(x) = |\{x_i \mid x_i \leq x\}|/n$$

Da die empirische Verteilungsfunktion eine Treppenfunktion ist (siehe Grafik am Ende des Kapitels), gibt es an jeder Stelle $x = x_i$ zwei mögliche Differenzen, die zu berücksichtigen sind, wenn wie im Folgenden das Supremum von $|F_{\text{emp}}(x) - F_0(x)|$ gesucht wird.

Es gilt

$$\begin{aligned}
 k &= \sup |F_{\text{emp}}(x) - F_0(x)| \\
 &= \max \left(\left\{ |F_{\text{emp}}(x_i) - F_0(x_i)| \mid i = 1, 2, \dots, n \right\} \right. \\
 &\quad \left. \cup \left\{ |F_{\text{emp}}(x_{i-1}) - F_0(x_i)| \mid i = 2, 3, \dots, n \right\} \cup \{F_0(x_0)\} \right)
 \end{aligned}$$

mit $F_0 = F_{N(\bar{x}, s^2)}$.

Die Prüfgröße ist definiert durch

$$w = \sqrt{n} \cdot k .$$

Die oben berechnete Prüfgröße w ist Realisierung einer Zufallsvariablen W , die (wie immer unter H_0) eine spezielle Verteilung besitzt. Diese Verteilung kann durch folgende Funktion approximiert werden:

$$F(x) = \begin{cases} \sum_{j=-\infty}^{\infty} (-1)^j \cdot e^{-2j^2 x^2} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

F ist die asymptotische Verteilung von W , unter der Voraussetzung, dass die beim Test verwendete Verteilungsfunktion F_0 keine unbekannt Parameter enthält die durch geschätzte Parameter ersetzt wurden.

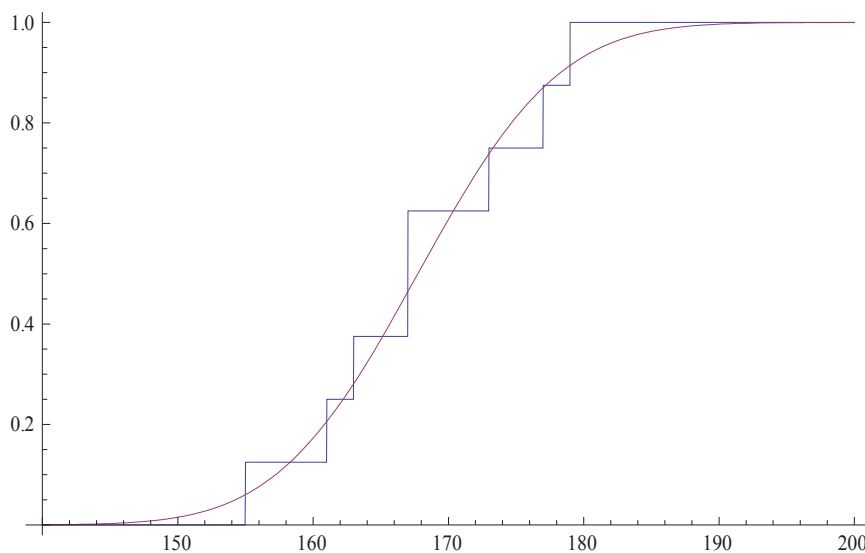
Im Beispiel ist $d \approx 0.16140$, $w \approx 0.456525$.

Für $n \leq 40$ können auch kritische Werte für k als Prüfgröße z.B. im Buch [3] gefunden werden, falls F_0 mit allen Parametern bekannt ist. Werden diese kritischen Werte bei einer Verteilung F_0 mit geschätzten Parametern verwendet (wie in unserem Fall), dann ist der Test

„konservativ“, d.h. dass damit H_0 seltener abgelehnt wird als man eigentlich mit exaktem p-Wert ablehnen würde.

Der approximative p-Wert $= 1 - F(w) \approx 0,98525 > 0,25$, womit die Nullhypothese nicht verworfen werden kann. Hier sollte man ein großes Signifikanzniveau verwenden, z.B. $\alpha = 25\%$, falls man auf der Basis der Normalverteilung weitere Tests durchführen möchte. Denn es handelt sich hierbei um einen Anpassungstest, man würde gerne H_0 zeigen. Da man aber den Fehler 2. Art nicht kennt (d.h. den Fehler, dass man H_0 nicht verwirft, obwohl H_0 falsch ist), ist hier ein großes Signifikanzniveau angebracht. Denn wenn man trotz eines großen Fehlers α , den man in Kauf nehmen würde, die Nullhypothese nicht verwerfen kann, dann spricht dies nicht gegen diese.

Es folgt noch eine Grafik, die F_{emp} und F_0 in unserem Beispiel zeigt. F_{emp} ist die Treppenfunktion, wobei die senkrecht eingezeichneten Linien nicht mit zur Funktion gehören.



Umsetzung mit SAS:

```
data dat1;  
input x;  
cards;  
167  
163  
155  
167  
161  
177  
173  
179  
run;  
  
proc univariate data = dat1 normal;  
var x;  
run;
```

SAS-Output zur Prozedur:

Die Prozedur UNIVARIATE
Variable: x

Momente			
N	8	Summe Gewichte	8
Mittelwert	167.75	Summe Beobacht.	1342
Std.abweichung	8.20713975	Varianz	67.3571429
Schiefe	-0.0441899	Kurtosis	-0.8864415
Unkorr. Qu.summe	225592	Korr. Quad.summe	471.5
Variationskoeff.	4.89248271	Stdfeh. Mittelw.	2.90166209

Grundlegende Statistikmaße			
Lage		Streuung	
Mittelwert	167.7500	Std.abweichung	8.20714
Median	167.0000	Varianz	67.35714
Modalwert	167.0000	Spannweite	24.00000
		Interquartilsabstand	13.00000

Tests auf Lageparameter: $\mu_0=0$				
Test	Statistik		p-Wert	
Studentsches t	t	57.81169	Pr > t 	<.0001
Vorzeichen	M	4	Pr >= M 	0.0078
Vorzeichen-Rang	S	18	Pr >= S 	0.0078

Tests auf Normalverteilung				
Test	Statistik		p-Wert	
Shapiro-Wilk	W	0.965247	Pr < W	0.8583
Kolmogorov-Smirnov	D	0.161406	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.028647	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.186586	Pr > A-Sq	>0.2500