Prof. Dr. Klaus Eckhardt

Chi-Quadrat-Verteilungstest

veröffentlicht im Internet unter aufgabomat.de

Gliederung:

1 Einleitung

Ziel der Statistik ist es, Zufallsvariablen zu beschreiben. Für eine Zufallsvariable sollen Kennwerte wie Erwartungswert oder Varianz ausgewiesen und Wahrscheinlichkeiten bestimmt werden können. Dies ist genau dann möglich, wenn die Wahrscheinlichkeitsverteilung bekannt ist, durch die sich die betreffende Variable beschreiben lässt.

Um diese Wahrscheinlichkeitsverteilung zu identifizieren, steht in der Praxis meist nur eine Stichprobe von Variablenwerten zur Verfügung. Man versucht dann beispielsweise durch die Bildung einer Häufigkeitsverteilung und aus ihrer grafischen Darstellung, dem Histogramm, zunächst auf den Typ der Wahrscheinlichkeitsverteilung zu schließen und berechnet anschließend aus den Stichprobenwerten Schätzwerte für die Parameter der Verteilung oder andere Kennwerte, etwa den empirischen Mittelwert als Schätzwert für den Erwartungswert.

Die so gefundene Hypothese bezüglich der beschreibenden Wahrscheinlichkeitsverteilung sollte vor weitergehenden Berechnungen allerdings abgesichert werden. Dazu dient die Gruppe der so genannten Verteilungs- oder Anpassungstests. Ein solcher ist der Chi-Quadrat-Test. Sich mit ihm zu befassen lohnt sich insofern besonders, da er nicht nur als Verteilungstest verwendet wird, sondern auch um Zusammenhänge zwischen Variablen zu identifizieren ($\rightarrow$ Kontingenzanalyse[1]).

2 Durchführung des Tests

Gegeben sei eine Stichprobe von N Werten einer Zufallsvariable X. Gesucht ist die Wahrscheinlichkeitsverteilung, durch die sich die Zufallsvariable beschreiben lässt. Bearbeitet wird diese Frage mit dem Chi-Quadrat-Verteilungstest.

1. Absolute Häufigkeit n_i der N Messwerte in I Klassen bzw. Intervallen ermitteln (i = 1, ..., I)

   Beispiel: Es sind N = 60 Werte des Körpergewichts von Frauen erfasst worden. Aus diesen Daten wird eine Häufigkeitsverteilung gebildet, die I = 9 Intervalle umfasst (Tabelle 1).

[1] Siehe Skript „Kontingenzanalyse“ unter aufgabomat.de.

i	Gewicht (kg)	empirische abs. Häufigkeit n_i
1	]35,0; 43,0]	1
2	]43,0; 51,0]	4
3	]51,0; 59,0]	6
4	]59,0; 67,0]	13
5	]67,0; 75,0]	16
6	]75,0; 83,0]	10
7	]83,0; 91,0]	7
8	]91,0; 99,0]	2
9	]99,0; 107,0]	1

Tabelle 1: Häufigkeitsverteilung einer Stichprobe von 60 Körpergewichtswerten.

2. Typ der Wahrscheinlichkeitsverteilung identifizieren
   Parameter der Wahrscheinlichkeitsverteilung abschätzen
   Null- und Alternativhypothese formulieren

   Im Allgemeinen fällt es leichter, den Typ der Wahrscheinlichkeitsverteilung aus einer grafischen Darstellung der Häufigkeitsverteilung, aus einem Histogramm, abzuleiten, als aus den Zahlen in einer Tabelle.

   Beispiel:

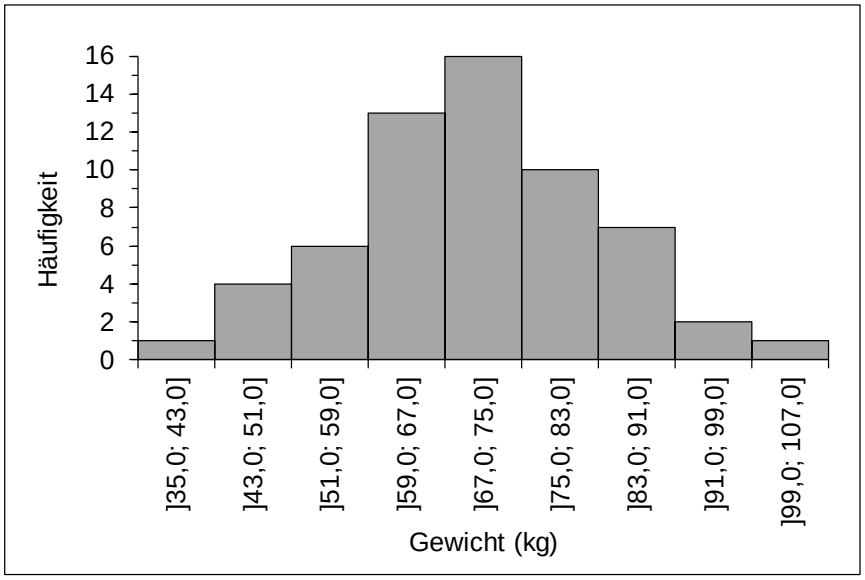

Abbildung 1: Histogramm zur Häufigkeitsverteilung in Tabelle 1.

Das Histogramm lässt auf eine eingipfelige Wahrscheinlichkeitsdichtefunktion schließen, d. h. auf eine Wahrscheinlichkeitsdichtefunktion mit nur einem Maximum, die symmetrisch zu diesem Maximum verlaufen könnte. Diese Eigenschaften weist die Wahrscheinlichkeitsdichtefunktion der Normalverteilung auf.

Die Normalverteilung hat zwei Parameter, den Mittelwert μ und die Standardabweichung σ. Schätzwert für den Mittelwert μ ist der empirische Mittelwert

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \, .$$

Schätzwert für die Standardabweichung σ ist die empirische Standardabweichung

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2} \; .$$

Im vorliegenden Beispiel ergibt sich $\bar{x}$ = 70,4 kg und s = 12,5 kg. Damit lassen sich als Null- und Alternativhypothese des Tests formulieren:

H_0: X normalverteilt mit μ = 70,4 kg und σ = 12,5 kg
H_1: X nicht normalverteilt mit μ = 70,4 kg und σ = 12,5 kg.

3. Ausgehend von der Nullhypothese hypothetische absolute Häufigkeiten n_i^* in den vorgegebenen I Klassen berechnen

Die zu erwartende Häufigkeit, mit der eine Zufallsvariable Werte in einem von mehreren Intervallen annimmt, berechnet sich, indem man die Wahrscheinlichkeit p_i für einen Wert in dem betreffenden Intervall mit der Gesamtzahl N der erfassten Werte multipliziert:

$$n_i^* = p_i \, N. \qquad (1)$$

Die Wahrscheinlichkeit p_i lässt sich im Fall stetiger Zufallsvariablen – beispielsweise dem Körpergewicht – am besten mit der Verteilungsfunktion der Wahrscheinlichkeitsverteilung berechnen. Diese wird im Folgenden mit F_0 bezeichnet. Es ist

$$p_i = F_0(\text{obere Begrenzung Intervall i}) - F_0(\text{untere Begrenzung Intervall i}). \qquad (2)$$

Beispiel: Hypothetische Wahrscheinlichkeitsverteilung ist die Normalverteilung. Werte der zugehörigen Verteilungsfunktion lassen sich in unterschiedlicher Weise gewinnen. Eine Variante besteht darin, die Intervallgrenzen einer Z-Transformation zu unterziehen und mit tabellierten Werten der Verteilungsfunktion der Standardnormalverteilung zu arbeiten[2]. Eine andere Variante ist die Verwendung eines Tabellenkalkulations- oder Statistikprogramms, in dem eine Funktion zur Berechnung der Verteilungsfunktionswerte implementiert ist. In Excel beispielsweise dient dazu die Funktion NORM.VERT.

Für i = 1 etwa ergibt sich

$$\begin{aligned} n_1^* &= [F_0(43{,}0) - F_0(35{,}0)] \cdot 60 \\ &= (0{,}014 - 0{,}002) \cdot 60 \\ &= 0{,}72 \end{aligned}$$

mit F_0: Verteilungsfunktion der Normalverteilung mit μ = 70,4 kg und σ = 12,5 kg.

Gerundet auf eine ganze Zahl ist $n_1^* = 1$. Nach analogen Berechnungen für die übrigen Gewichtsintervalle ergibt sich die in der rechten Spalte der Tabelle 2 aufgeführte hypothetische Häufigkeitsverteilung.

[2] Siehe Skript „Standardisierung der Normalverteilung“ unter aufgabomat.de.

i	Gewicht (kg)	empirische abs. Häufigkeit n_i	hypothetische abs. Häufigkeit n_i^*
1	]35,0; 43,0]	1	1
2	]43,0; 51,0]	4	3
3	]51,0; 59,0]	6	7
4	]59,0; 67,0]	13	13
5	]67,0; 75,0]	16	15
6	]75,0; 83,0]	10	12
7	]83,0; 91,0]	7	6
8	]91,0; 99,0]	2	2
9	]99,0; 107,0]	1	1

Tabelle 2: Hypothetische Häufigkeit im Beispiel des Körpergewichts.

4. Klassen so zusammenfassen, dass n_i^* überall mindestens 5 beträgt

   Die reduzierte Anzahl der Klassen wird mit I^* bezeichnet.

   Beispiel: In den Intervallen bzw. Klassen 1 und 2 ist die hypothetische Häufigkeit jeweils kleiner als 5. Auch wenn diese beiden Klassen zusammengefasst werden, beträgt die hypothetische Häufigkeit lediglich 4. Daher werden hier sogar die ersten drei Intervalle vereint. Ebenso müssen die letzten drei Intervalle zusammengefasst werden. Es ergibt sich $I^* = 5$.

i	Gewicht (kg)	empirische abs. Häufigkeit n_i	hypothetische abs. Häufigkeit n_i^*
1	]35,0; 59,0]	11	11
2	]59,0; 67,0]	13	13
3	]67,0; 75,0]	16	15
4	]75,0; 83,0]	10	12
5	]83,0; 107,0]	10	9

Tabelle 3: Empirische und hypothetische Häufigkeit.

5. Teststatistik berechnen
   Gültigkeit der Nullhypothese beurteilen

   Nun wird der Wert einer Variable, der so genannten Teststatistik, berechnet, die dann, wenn die Nullhypothese gilt, bekannte Eigenschaften aufweist. Der Formulierung der Teststatistik liegen die folgenden Gedanken zugrunde:
   - Je ähnlicher n_i und n_i^* sind, d. h. je kleiner die Differenzen $n_i - n_i^*$, desto wahrscheinlicher gilt die Nullhypothese.
   - Um zu verhindern, dass sich negative und positive Differenzen gegenseitig aufheben, werden die Differenzen quadriert.
   - Eine Differenz fällt umso weniger ins Gewicht, je mehr Werte ohnehin in der jeweiligen Klasse zu erwarten sind. Daher werden die quadrierten Differenzen noch durch n_i^* geteilt.

   Die Teststatistik des Chi-Quadrat-Verteilungstests ist

$$\chi^2 = \sum_{i=1}^{I^*} \frac{(n_i - n_i^*)^2}{n_i^*} . \qquad (3)$$

Das Symbol χ ist der griechische Großbuchstabe Chi.

Falls die Nullhypothese H_0 zutrifft und Anzahl der Messwerte genügend groß ist (Faustregel: N > 50), ist χ^2 Chi-Quadrat-verteilt. Aus Gleichung 3 ist ersichtlich, dass Chi-Quadrat-verteilte Variablen nur Werte ≥ 0 annehmen. Die Wahrscheinlichkeitsdichtefunktion der Chi-Quadrat-Verteilung ist daher nach links durch die Null begrenzt, nach rechts aber unbegrenzt und damit asymmetrisch. Entsprechendes gilt für die zugehörige Verteilungsfunktion (Abbildung 2).

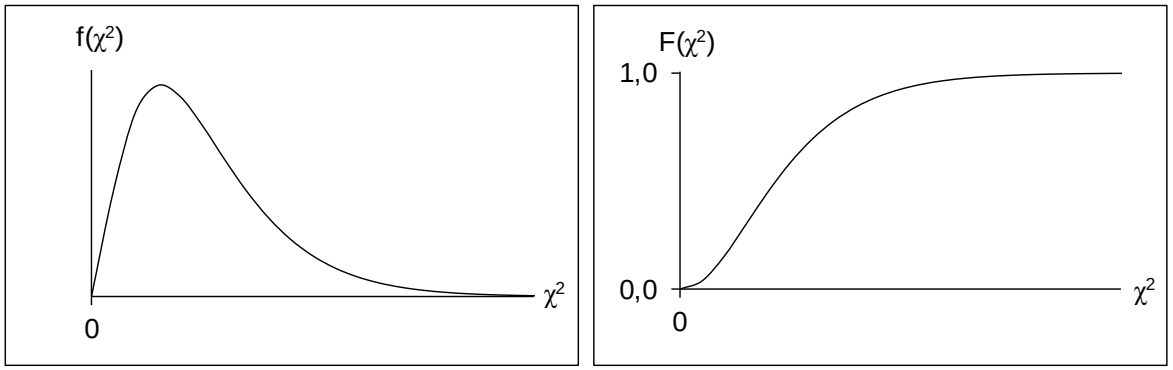

Abbildung 2: Wahrscheinlichkeitsdichte- und Verteilungsfunktion der Chi-Quadrat-Verteilung.

Die Chi-Quadrat-Verteilung hat nur einen Parameter, den so genannten Freiheitsgrad, für den hier symbolisch f geschrieben wird. Im Fall des Verteilungstests ist

f = I* – 1 – Anzahl der empirisch bestimmten Parameter der hypothetischen Verteilung. (4)

Beispiel:

i	Gewicht (kg)	empirische abs. Häufigkeit n_i	hypothetische abs. Häufigkeit n_i*	$n_i - n_i$*	$(n_i - n_i^*)^2/n_i^*$
1	]35,0; 59,0]	11	11	0	0,0
2	]59,0; 67,0]	13	13	0	0,0
3	]67,0; 75,0]	16	15	1	0,1
4	]75,0; 83,0]	10	12	–2	0,3
5	]83,0; 107,0]	10	9	1	0,1

Tabelle 4: Berechnung des Werts der Teststatistik.

$\chi^2 = 0{,}5$

Je kleiner der Wert der Teststatistik ist, desto wahrscheinlicher ist es, dass die Nullhypothese zutrifft. $\chi^2 = 0$ bedeutet, dass keinerlei Unterschied zwischen empirischer und hypothetischer Häufigkeitsverteilung besteht. Der Chi-Quadrat-Verteilungstest wird daher rechtsseitig durchgeführt, d. h. das Annahmeintervall wird nur nach rechts durch ein Quantil der Chi-Quadrat-Verteilung begrenzt.

Beispiel: Angenommen, der Chi-Quadrat-Test wird durchgeführt, um die Voraussetzung der Normalverteilung bei weitergehenden Analyseverfahren zu prüfen. In diesem Fall ist es besonders kritisch, wenn die Nullhypothese irrtümlich beibehalten wird. Die Wahrscheinlichkeit für den Fehler zweiter Art sollte daher klein sein. Die einzige Möglichkeit, dies zu erreichen, ist, die Wahrscheinlichkeit für den Fehler erster Art, die Irrtumswahrscheinlichkeit α, groß zu wählen. Daher wird hier nicht mit dem meist verwendeten α = 0,05, sondern mit α = 0,10 gerechnet.

Als rechtsseitige Begrenzung des Annahmeintervalls ist das 0,90-Quantil $\chi^2_{0,90}$ der Chi-Quadrat-Verteilung mit f = 5 – 1 – 2 = 2 zu ermitteln, denn zwei Parameter der hypothetischen Verteilung, μ und σ, sind empirisch bestimmt worden (→ Schritt 2). Wie sich beispielsweise einer Quantiltabelle entnehmen lässt (Tabelle 5), ist $\chi^2_{0,90} = 4,6$.

f	p		
	0,90	0,95	0,99
1	2,7	3,8	6,6
2	4,6	6,0	9,2
3	6,3	7,8	11,3
4	7,8	9,5	13,3
5	9,2	11,1	15,1
6	10,6	12,6	16,8
7	12,0	14,1	18,5
8	13,4	15,5	20,1
9	14,7	16,9	21,7
10	16,0	18,3	23,2

f	p		
	0,90	0,95	0,99
11	17,3	19,7	24,7
12	18,5	21,0	26,2
13	19,8	22,4	27,7
14	21,1	23,7	29,1
15	22,3	25,0	30,6
16	23,5	26,3	32,0
17	24,8	27,6	33,4
18	26,0	28,9	34,8
19	27,2	30,1	36,2
20	28,4	31,4	37,6

Tabelle 5: Quantile χ^2_p der Chi-Quadrat-Verteilung.

Da $\chi^2 \leq \chi^2_{0,90}$, wird H_0 beibehalten. Es kann davon ausgegangen werden, dass die Daten normalverteilt mit μ = 70,4 kg und σ = 12,5 kg sind.

Die Eingangsdaten des in diesem Skript behandelten Beispiels und weitere Übungsaufgaben finden sich im Internet unter der Adresse aufgabomat.de in der Rubrik Statistik.