Lösung zu Aufgabe 11: Chi²-Test

Häufig wird bei der Bearbeitung statistischer Daten eine bestimmte Verteilung vorausgesetzt. Um zu überprüfen ob die Daten tatsächlich der Verteilung entsprechen, wird ein χ^2*-Test durchgeführt. Dabei sei* X *eine Zufallsgröße mit unbekannter Verteilungsdichtefunktion. Aufgrund von Messdaten oder Vorabinformationen wird vermutet, dass* X *durch eine Verteilungsdichtefunktion* $h(x)$ *beschrieben wird.*

Die hierzu formulierte Nullhypothese H_0 *lautet:* X *wird durch die Verteilungsdichtefunktion* $h(x)$ *beschrieben! Etwaig beobachtete Abweichungen von der angenommenen Verteilung wären in diesem Fall rein zufälliger Natur.*

Weiterhin wird eine Stichprobe mit n *Messwerten* $x_1,...,x_n$ *aufgenommen.*

Aus dieser Messreihe wird einerseits ein empirisches Histogramm erstellt und aus der Verteilungsdichtefunktion $h(x)$ *wird ein theoretisches Histogramm berechnet.*

Als Testgröße wird eine normierte Differenz zwischen beiden Histogrammen berechnet. Wenn die Hypothese zutrifft, müsste diese Testgröße hinreichend klein sein.

Anhand nachfolgender Abbildungen soll die grundsätzliche Vorgehensweise zunächst nochmals anschaulich dargestellt werden, bevor im zweiten Teil dieses Dokuments die rechnerische Durchführung des χ^2-Tests erfolgt.

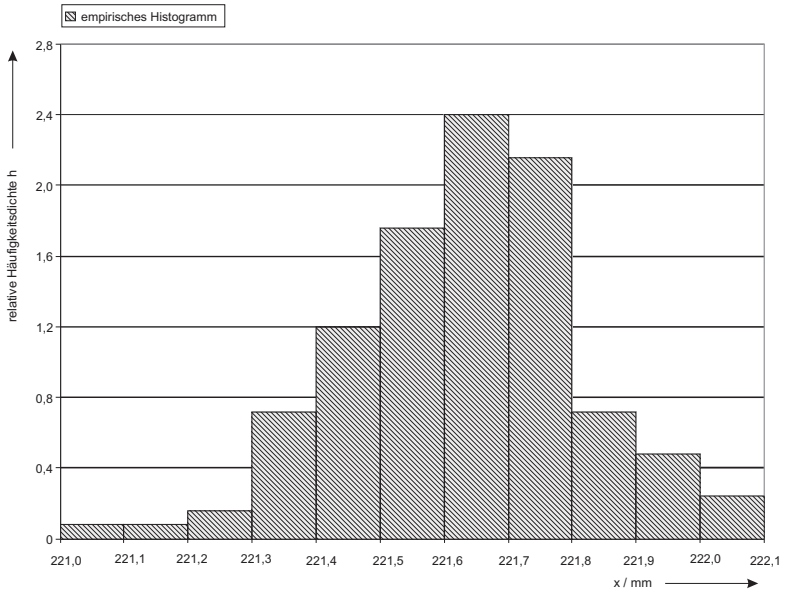

Abbildung 1: Empirisches Histogramm der Messdaten

Im vorliegenden Fall der Übungsaufgabe 11 wurden die aufgenommenen Messwerte bereits zu einem Histogramm zusammen gefasst. Die entsprechenden Klassengrenzen sowie die absoluten Häufigkeiten in den einzelnen Klassen können der Tabelle in der Aufgabenstellung entnommen werden. Abbildung 1 zeigt dieses empirische Histogramm, wobei hier wie auch in den nachfolgenden Abbildungen jeweils die Fläche der Histogrammbalken der relativen

Häufigkeit proportional ist. (Das heißt, die absoluten Häufigkeiten lassen sich aus der Balkenhöhe durch Multiplikation mit dem Stichprobenumfang von n = 125 sowie mit der Klassenbreite von 0,1 mm errechnen.)

Das in Abbildung 1 dargestellt Histogramm weist zunächst noch Klassen einheitlicher Breite auf. Da für dünn besetzte Klassen die oben erwähnte Berechnung der *normierten* Differenzen von empirischem und theoretischem Histogramm zu unsinnig großen Zahlenwerten führen kann, sollten die Histogrammklassen so gewählt werden, dass in jede Klasse mindestens 5 Messwerte entfallen. Bei diesem Zahlenwert handelt es sich allerdings nur um einen Richtwert, man findet mitunter auch die Empfehlung, dass jede Klasse mit mindestens 10 Messwerten besetzt sein sollte. Am einfachsten lässt sich diese Bedingung dadurch einhalten, dass in dem bereits vorliegenden Histogramm solche dünn besetzten Klassen mit benachbarten Klassen zusammen gelegt werden, bis eine entsprechende Mindestbesetzung erreicht ist. Im vorliegenden Fall werden daher die ersten vier Klassen zu einer dann von 221,0 mm bis 221,4 mm reichenden Klasse zusammen gefasst, sowie die beiden letzten Klassen zu einer dann von 221,9 mm bis 222,1 mm reichenden. Die grafische Darstellung dieses empirischen Histogramms mit zusammen gefassten Klassen findet sich in Abbildung 2.

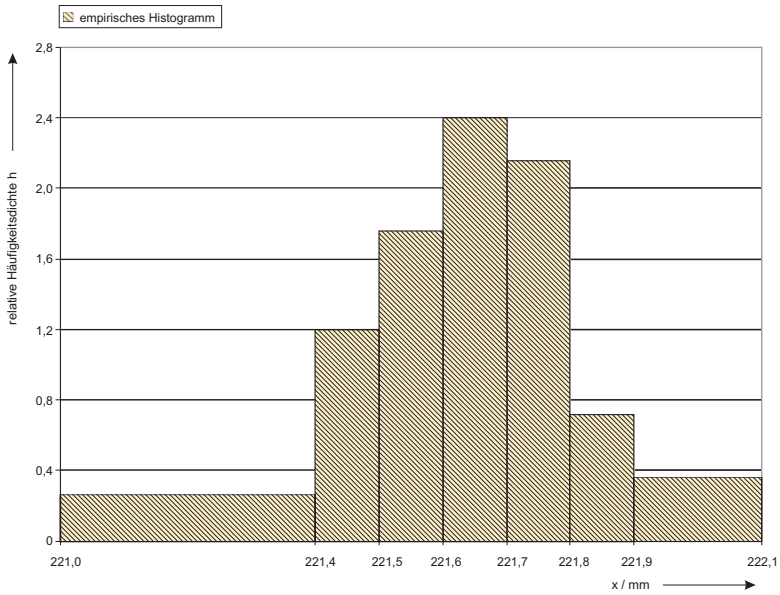

Abbildung 2: Empirisches Histogramm der Messdaten mit zusammen gefassten Klassen

Durch die Aufgabenstellung ist vorgegeben, dass als Testhypothese von einer Gaußschen Normalverteilung ausgegangen werden soll. Die Parameter μ und σ der am besten zu den vorliegenden Messdaten passende Normalverteilung werden durch den Mittelwert $\bar{x}$ und die Streuung S der Messdaten abgeschätzt. In Abbildung 3 ist zur Veranschaulichung die Verteilungsdichtefunktion der so ermittelten Normalverteilung dem empirischen Histogramm der Messdaten überlagert.

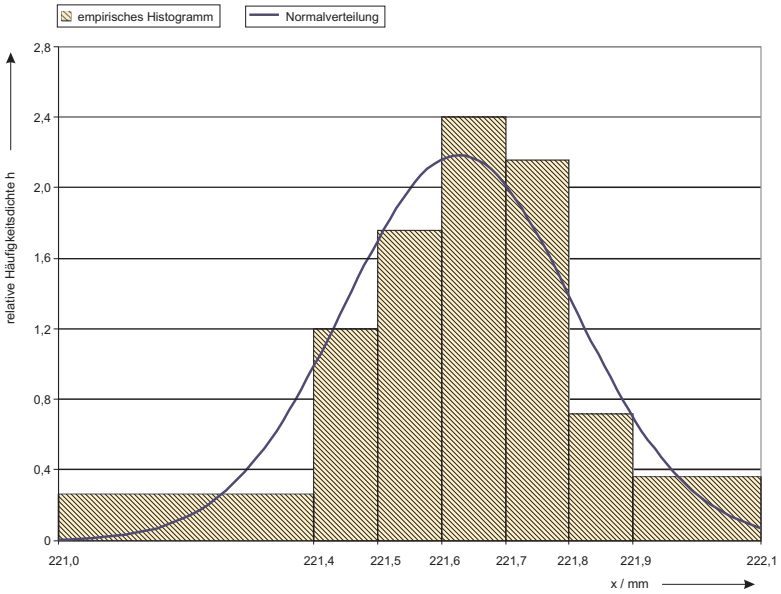

Abbildung 3: Empirisches Histogramm mit daraus abgeleiteter Normalverteilung

Im nächsten Schritt wird nun unter Zugrundelegung der zuvor abgeschätzten Normalverteilung ein theoretisches Histogramm ermittelt. Das heißt, es wird berechnet, wie viele Messwerte theoretisch in die zuvor gebildeten Histogrammklassen entfallen würden, wenn man davon ausgeht, dass der Verteilung *tatsächlich* die ermittelte Normalverteilung zugrunde liegt. In Abbildung 4 ist das resultierende theoretische Histogramm gemeinsam mit der zugrunde liegenden Normalverteilung dargestellt.

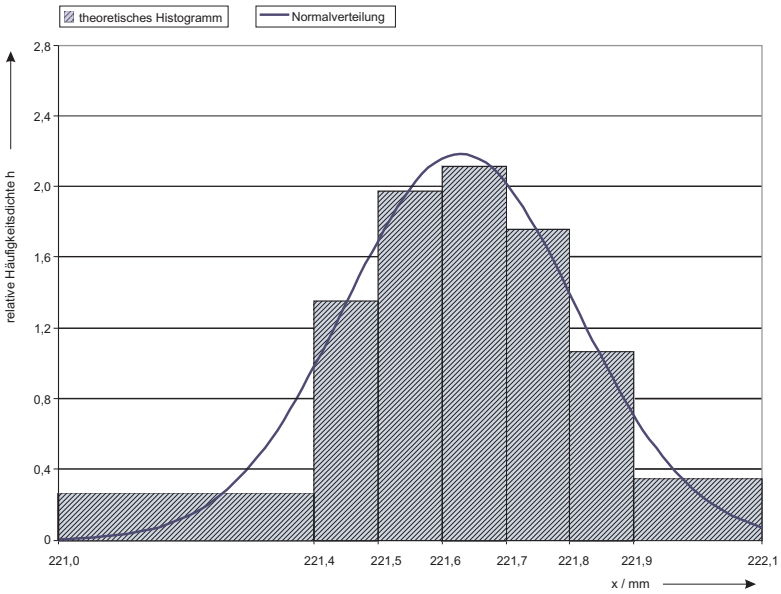

Abbildung 4: Theoretisches Histogramm mit zugrunde liegender Normalverteilung

Vergleich man nun das empirische Histogramm aus Abbildung 2 mit dem theoretischen Histogramm aus Abbildung 4, stellt man fest, dass sich zwischen tatsächlicher Besetzungszahl der Klassen und theoretisch erwarteter Besetzungszahl Unterschiede ergeben. Eine grafische Darstellung dieser Gegenüberstellung von empirischem und theoretischem Histogramm zeigt Abbildung 5.

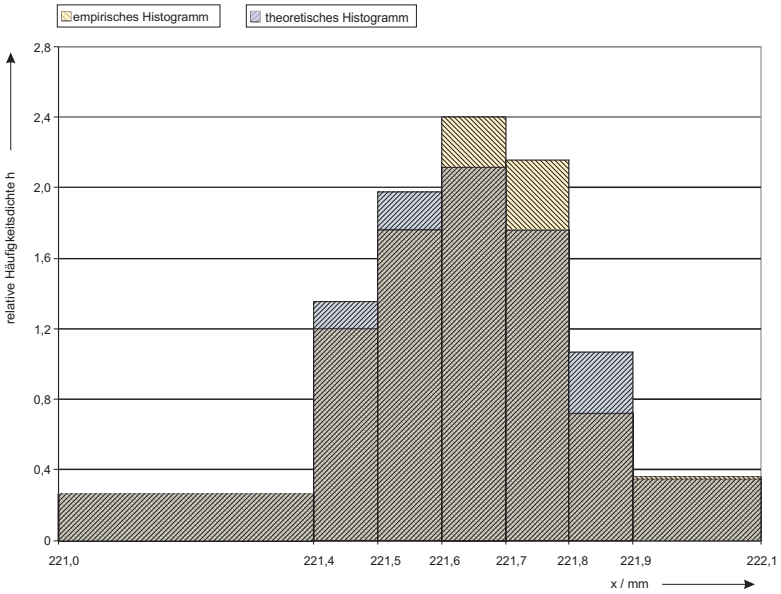

Abbildung 5: Vergleich von empirischem und theoretischem Histogramm

Die zu klärende Frage ist nun, ob die beobachteten Unterschiede so gering sind, dass man mit hinreichender statistischer Sicherheit davon ausgehen kann, dass sie lediglich zufälliger Natur sind, oder ob die Unterschiede so groß sind, dass die Hypothese von der Normalverteilung der erhobenen Messdaten verworfen werden muss. Hierzu wird aus den Differenzen der beiden Histogramme eine normierte Gesamtdifferenz χ_0^2 berechnet und diese mit einem kritischen Wert verglichen. Ist die errechnete Differenz kleiner als der kritische Wert, wird die Nullhypothese, die Messdaten seien normalverteilt, angenommen, andernfalls wird sie verworfen.

Zur rechnerischen Bestimmung des χ_0^2-Wertes dient untenstehende Tabelle. In der ersten Spalte dieser Tabelle sind die Obergrenzen x_i der in der Aufgabenstellung gegebenen Histogrammklassen eingetragen, in der zweiten Spalte finden sich die absoluten Häufigkeiten n_i innerhalb dieser Klassen.

Zunächst werden nun wie oben beschrieben dünn besetzte Klassen mit benachbarten Klassen zusammen gefasst. Wie zu erkennen, sind die ersten drei Klassen nur mit Häufigkeiten von 1 bzw. 2 besetzt. Da auch nach Zusammenfassung dieser drei benachbarten Klassen noch nicht die Mindestanzahl von 5 erreicht ist, müssen wir die ersten vier Klassen zusammen fassen. Ebenso muss die letzte Klasse (Besetzungszahl 3) mit der vorletzten Klasse zusammen gefasst werden. Die empirisch ermittelten absoluten Häufigkeiten B_i innerhalb der verbleibenden Klassen sind in der dritten Spalte der Tabelle eingetragen.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
x_i	n_i	B_i	$(x_i-\mu)/\sigma$	$\Phi(x_i-\mu)/\sigma$	p_i	$E_i = np_i$	$\frac{(B_i-E_i)^2}{E_i}$
221,1	1	13	–1,26	0,103835	0,103835	12,9794	0,00003
221,2	1						
221,3	2						
221,4	9						
221,5	15	15	–0,71	0,238852	0,135017	16,8771	0,2088
221,6	22	22	–0,16	0,436441	0,197589	24,6986	0,2949
221,7	30	30	0,38	0,648027	0,211586	26,4483	0,4770
221,8	27	27	0,93	0,823814	0,175787	21,9734	1,1499
221,9	9	9	1,48	0,930563	0,106749	13,3436	1,4139
222,0	6	9	∞	1,000000	0,069437	8,6796	0,0118
∞	3						
						χ_0^2	**3,5563**

Die Berechnung des theoretischen Histogramms erfolgt im vorliegenden Fall unter Zuhilfenahme der bereits aus vorangegangenen Aufgaben bekannten Tabelle der Summenfunktion der standardisierten Normalverteilung. Dazu berechnen wir zunächst aus den x_i-Werten der Klassenobergrenzen die korrespondierenden, auf die standardisierte Normalverteilung bezogenen z-Koordinaten. Hierfür benötigen wir jedoch zunächst den Erwartungswert μ und die Standardabweichung σ der anzunehmenden Normalverteilung. Die besten Schätzwerte für diese beiden Parameter stellen der Mittelwert $\bar{x}$ und die Streuung S der vorliegenden Messdaten dar. Beide Werte sind bereits in der Aufgabenstellung angegeben und lauten:

$\bar{x} = 221{,}63 \text{ mm}$

$s = 0{,}183 \text{ mm}$

Für die erste Klasse mit der Obergrenzen von $x_1 = 221{,}4 \text{ mm}$ lautet die Berechnung des z-Wertes daher:

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{221{,}4 \text{ mm} - 221{,}63 \text{ mm}}{0{,}183 \text{ mm}} \approx -1{,}26$$

Dieser Zahlenwert, sowie die analog hierzu berechneten Werte für die folgenden Klassen finden sich in der vierten Spalte der Tabelle. Für die letzte Klasse, deren Obergrenze hier mit ∞ gewählt wurde, ergibt sich ohne Berechnung auch für den z-Wert ∞.

Die zu den berechneten z-Werten gehörigen Werte der Summenfunktion der standardisierten Normalverteilung $\Phi(z_i)$ können nun nach bekanntem Schema aus der entsprechenden Tabelle

abgelesen werden. Hierdurch erhält man die in der fünften Spalte eingetragenen Zahlenwerte. Der Wert in der letzten Zeile, für $z = \infty$, kann zwar nicht aus der Tabelle abgelesen werden, ergibt sich jedoch logischer Weise zu 1, da zwischen $-\infty$ und $+\infty$ stets 100% aller Messwerte liegen.

Die bislang ermittelten, in Spalte 5 eingetragenen Werte geben nun an, mit welcher Wahrscheinlichkeit ein Wert zwischen $-\infty$ und der jeweiligen Klassenobergrenze liegt. Da wir jedoch an der Wahrscheinlichkeit interessiert sind, mit welcher ein Wert *innerhalb* der Klassengrenzen liegt, müssen wir die entsprechenden Differenzen bilden. Für die erste Klasse können wir den Wert von $\Phi(z_1) = 0{,}103835$ unverändert übernehmen. Für die zweite Klasse erhalten wir das Ergebnis, indem wir $\Phi(z_1)$ von $\Phi(z_2)$ subtrahieren. Wir erhalten also:

$$\Phi(z_2) - \Phi(z_1) = 0{,}238852 - 0{,}103835 = 0{,}135017$$

Die nach diesem Schema errechneten und in Spalte 6 der Tabelle eingetragenen Werte geben nun an, mit welcher Wahrscheinlichkeit p_i unter Annahme der geschätzten Normalverteilung ein Messwert innerhalb einer bestimmten Klasse liegt. Da unser empirisches Histogramm die absoluten Häufigkeiten B_i innerhalb der Klassen angibt, müssen wir für den Vergleich von empirischem und theoretischem Histogramm die Wahrscheinlichkeiten p_i aus Spalte 6 in absolute Häufigkeiten E_i umrechnen. Hierzu multiplizieren wir die Wahrscheinlichkeiten mit dem Umfang der Stichprobe, welche den empirischen Daten zugrunde liegt. Dieser Stichprobenumfang beträgt $n = 125$. So errechnen wir beispielsweise die absolute Häufigkeit innerhalb der ersten Klasse des theoretischen Histogramms wie folgt:

$$E_1 = n \cdot p_1 = 125 \cdot 0{,}103835 \approx 12{,}9794$$

Bevor wir nun mit Hilfe der 3. und 7. Spalte der Tabelle die normierte Differenz von empirischem und theoretischem Histogramm berechnen, überprüfen wir, ob auch für das theoretische Histogramm die Bedingung erfüllt ist, dass die absolute Häufigkeit in keiner der Klassen den Wert 5 unterschreitet. Sollte dies für einzelne Klassen der Fall sein, fassen wir – wie bereits anhand des empirischen Histogramms erläutert – benachbarte Klassen zusammen. Im vorliegenden Fall ist jedoch keine weitere Zusammenlegung von Klassen erforderlich.

Die normierte Differenz χ_0^2 errechnet sich nun aus den empirischen Häufigkeiten B_i und den theoretischen Häufigkeiten E_i gemäß folgender Gleichung:

$$\chi_0^2 = \sum_{i=1}^{r^*} \frac{(B_i - E_i)^2}{E_i}$$

Wir errechnen hier zunächst die normierten Differenzen jeweils für einzelne Klassen und erhalten damit die in Spalte 8 der Tabelle eingetragenen Zahlenwerte. Für die erste Klasse lautet die Berechnung beispielsweise:

$$\frac{(B_1 - E_1)^2}{E_1} = \frac{(13 - 12{,}9794)^2}{12{,}9794} \approx 3{,}27 \cdot 10^{-5}$$

Bilden wir nun über diese klassenweisen normierten Differenzen die Summe, erhalten wir den gesuchten χ_0^2-Wert für den damit gilt:

$$\chi_0^2 = 3{,}5563$$

Die im Weiteren auszuwertende Testbedingung lautet:

$$\chi_0^2 > \chi^2_{r^*-s-1;1-\alpha}$$

Wir müssen demnach zunächst den kritischen Wert $\chi^2_{r^*-s-1;1-\alpha}$ ermitteln. Hierfür benötigen wir neben dem gewählten Signifikanzniveau α die beiden Parameter r^* und s. r^* steht hierbei für die Anzahl der auswertbaren Klassen des Histogramms. Auswertbar meint hierbei, die Zahl der Klassen nach einer etwaigen Zusammenlegung benachbarter Klassen. Im vorliegenden Fall erhalten wir durch Auszählen:

$$r^* = 7$$

Der Parameter s steht für die Anzahl der aus der untersuchten Stichprobe abgeschätzten Parameter der Verteilungsdichtefunktion, die dem jeweiligen Test zugrunde gelegt wird. Im vorliegenden Fall war dies die Verteilungsdichtefunktion der Gaußschen Normalverteilung. Diese lautet:

$$h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Die Verteilungsdichtefunktion der Gaußschen Normalverteilung hängt also von den zwei Parametern μ und σ ab. Beide Werte wurden in Form des Mittelwerts $\bar{x}$ bzw. der Streuung S aus den Daten der untersuchten Stichprobe abgeschätzt. Die gesuchte Anzahl s beträgt daher:

$$s = 2$$

Das Signifikanzniveau α beträgt laut Aufgabenstellung $\alpha = 0{,}05$. Für den kritischen Wert $\chi^2_{r^*-s-1;1-\alpha}$ gilt daher im vorliegenden Fall:

$$\chi^2_{r^*-s-1;1-\alpha} = \chi^2_{7-2-1;1-0{,}05} = \chi^2_{4;0{,}95}$$

Der zugehörige Zahlenwert kann nun aus der Tabelle der p-Quantile $\chi^2_{s;p}$ der χ^2-Verteilung mit s Freiheitsgraden abgelesen werden. Wir erhalten somit:

$$\chi^2_{4;0{,}95} = 9{,}49$$

Die auszuwertende Testbedingung lautet mit den errechneten Zahlenwerten folglich:

$$3{,}5563 > 9{,}49$$

Wie zu erkennen, ist diese Testbedingung nicht erfüllt. Da für denn Fall, dass die Testbedingung erfüllt ist, die Nullhypothese abgelehnt wird, schließen wir aus der Nichterfüllung der Testbedingung:

Die Nullhypothese wird *nicht* abgelehnt!

Da unsere Nullhypothese lautet, dass die vorliegenden Messdaten einer Gaußschen Normalverteilung genügen, schließen wir aus dem Testergebnis weiterhin:

Die Messdaten sind mit einer statistischen Sicherheit von P = 95% normalverteilt!