

statistischen Test zunächst nur den Fehler 1. Art kontrollieren können, dass nämlich die Alternativhypothese akzeptiert wird, obwohl sie falsch ist. Also ist der χ^2 -Anpassungstest eigentlich ein Test darauf, dass eine Zufallsvariable einer bestimmten Verteilung *nicht* folgt, als dass sie dieser Verteilung folgt. Wir können den Fehler 2. Art, dass wir die Nullhypothese akzeptieren, obwohl sie falsch ist, nicht quantifizieren.

Die Nullhypothese soll nun anhand einer Stichprobe überprüft werden. Wir nehmen an, dass in einer Stichprobe vom Umfang n der Wert y_i gerade n_i -mal vorkommt, mit $n_1 + \dots + n_r = n$. Diese n_i sind gerade die absoluten Häufigkeiten, wie man sie auch in Säulendiagrammen verwendet. Im Kontext des Anpassungstests nennt man sie auch **beobachtete Häufigkeiten**. Nehmen wir nun an, dass die Nullhypothese wahr ist, so finden wir mit Wahrscheinlichkeit p_i beim zufälligen Herausgreifen eines Objekts aus der Grundgesamtheit den Wert y_i . Der Erwartungswert für N_i (die zu n_i gehörige Zufallsvariable) ist somit gleich $e_i = np_i$. Diese Zahlen e_i nennen wir die **erwarteten Häufigkeiten** für die Werte y_i . Als Maß für die Abweichung der Stichprobe von der hypothetischen Verteilung dient die Summe der *relativen quadratischen Abweichungen* der beobachteten von den erwarteten Häufigkeiten:

$$\chi^2 := \sum_{i=1}^r \frac{(n_i - e_i)^2}{e_i}.$$

Man kann zeigen, dass die zugehörige Zufallsvariable

$$X^2 := \sum_{i=1}^r \frac{(N_i - e_i)^2}{e_i},$$

für große n näherungsweise χ^2 -verteilt ist mit $k = r - 1$ Freiheitsgraden, d.h.

$$\mathbb{P}(X^2 < x^2) = c \int_0^{x^2} u^{k/2-1} e^{-u/2} du. \quad (6.21)$$

Dabei ist c eine explizit bekannte Normierungskonstante. Diese Dichte verschwindet für negative Werte von X^2 – schließlich können Abweichungsquadraten nie negativ sein. Der Test ist also immer *einseitig*. Die Nullhypothese ist also abzulehnen, falls für eine vorgegebene Irrtumswahrscheinlichkeit α die Ungleichung

$$\chi^2 > \chi_{k,1-\alpha}^2$$

gilt, wobei $\chi_{k,q}^2$ das q -Quantil der χ^2 -Verteilung mit k Freiheitsgraden bezeichnet, also denjenigen Wert, der

$$\mathbb{P}(X^2 \leq \chi_{k,q}^2) = q$$

erfüllt. (Entsprechende Quantilwerte entnimmt man der χ^2 -Verteilungstabelle).

Bemerkung. Die Symbole X^2 , χ^2 und x^2 haben sich eingebürgert, obwohl sie leicht missverständlich sind. Der Exponent, der an den Ursprung als Quadratsumme erinnert, ist Teil des Symbols; X^2 , χ^2 und x^2 werden also wie gewöhnliche Variablen behandelt. X^2 ist nicht mit dem Quadrat einer Zufallsvariablen X zu verwechseln!

Es ist zu beachten, dass die χ^2 -Verteilung eine schlechte Näherung ist, wenn die Stichprobe zu klein ist. Als Faustregel gilt: Der Test darf im Fall $r \leq 8$ nur benutzt werden, wenn alle $e_i \geq 5$ sind, und im Fall $r > 8$, wenn alle $e_i \geq 1$. Wenn die Daten (mindestens) ordinalskaliert sind und n groß genug ist, lassen sich die Faustregeln erfüllen, indem man benachbarte Werte zu größeren Klassen zusammenfasst.

Beispiel. Wir betrachten eine Pflanzenart mit den Genotypen aa (1; weiße Blüte), Aa (2; rosa Blüte) und AA (3; rote Blüte) und kreuzen Aa mit Aa. Falls der intermediäre Mendelsche Erbgang die Situation perfekt beschreibt, erwartet man in der Nachkommenschaft aa, Aa und AA im Verhältnis $p_1 = 1/4$, $p_2 = 1/2$, und $p_3 = 1/4$. Anhand einer Stichprobe der Größe $n = 400$ soll überprüft werden, ob die Verteilung der Nachkommenschaft mit den Mendelschen Verhältnissen kompatibel ist. Die Stichprobe liefert die beobachteten Häufigkeiten $n_1 = 115$, $n_2 = 171$, $n_3 = 114$. Die unter H_0 erwarteten Häufigkeiten sind $e_1 = 100$, $e_2 = 200$, $e_3 = 100$. Man erhält $\chi^2 = (115-100)^2/100 + (171-200)^2/200 + (114-100)^2/100 = 8.46 > 5.99 = \chi^2_{2,0.95}$, H_0 wird also abgelehnt. (Nun darf man nach den biologischen Ursachen für die Abweichung von den Mendelschen Verhältnissen forschen: z.B. könnte Selektion gegen Heterozygote im Spiel sein, oder Wechselwirkung des Gens für die Blütenfarbe mit anderen Teilen des Genoms.)

Bisher haben wir nur die Situation betrachtet, dass wir die Verteilung unter H_0 vollständig kennen, ohne dass irgendwelche Parameter spezifiziert werden mussten. Oft steht man aber vor der Situation, dass man einen (oder mehrere) Parameter der hypothetischen Verteilung erst aus der Stichprobe schätzen muss. Dann verringert sich die Zahl der Freiheitsgrade um a , die Zahl der geschätzten Parameter, also $k = r - a - 1$.

Beispiel. Wir betrachten die Zahl der radioaktiven Zerfälle pro Zeiteinheit (wie in einer der Hausaufgaben). Es soll zum Signifikanzniveau $\alpha = 0.05$ überprüft werden, ob die Beobachtungen mit einer Poisson-Verteilung verträglich sind. Dazu muss zunächst der Parameter der Poisson-Verteilung aus der Stichprobe geschätzt werden. Es ergibt sich $\hat{\lambda} = \bar{y} = 5.4$. Mit Hilfe dieser Schätzung werden die erwarteten Häufigkeiten ermittelt: $e_i = n \cdot \mathbb{P}(Y = i) = n \frac{\hat{\lambda}^i}{i!} e^{-\hat{\lambda}}$. Die Beobachtungen werden so zu Klassen zusammengefasst, dass alle e_i größer als 1 sind.

Anz. Zerfälle	≤ 1	2	3	4	5	6	7	8	9	10	≥ 11
beobachtet	1	2	10	9	6	5	7	3	5	1	1
erwartet	0.23+1.22 = 1.45	3.29	5.93	8.00	8.64	7.78	6.00	4.05	2.43	1.31	0.64+0.29+... = 1.12

Wir haben $r = 11$ Klassen, die Zahl der Freiheitsgrade beträgt $r - 1 - 1 = 9$, und es gilt $\chi^2 = 8.61 < \chi_{9,0.95}^2 = 16.92$. Also ist keine signifikante Abweichung von der Poissonverteilung festzustellen.

6.3.3 Der Chi-Quadrat-Anpassungstests im allgemeinen Fall

Der χ^2 -Anpassungstest aus dem letzten Abschnitt eignet sich für diskrete Zufallsvariablen mit endlichen vielen Realisierungen. Wir behandeln hier den allgemeinen Fall, welcher auch diskrete Zufallsvariablen mit unendlich vielen Realisierungen sowie kontinuierliche Zufallsvariablen einschließt. Dabei werden wir den allgemeinen Fall durch Klassenbildung auf den endlichen Fall zurückführen.

Sei F die unbekannte (nicht notwendig diskrete) Verteilungsfunktion der reellwertigen Zufallsvariablen X . Wir betrachten die jeweilige Nullhypothese für den Fall einer vollständig bekannten zu überprüfenden Verteilung, sowie für den Fall einer von unbekannten, noch zu schätzenden Parametern abhängigen Verteilung:

- (a) $H_0 : F = F_0$; $H_A : F \neq F_0$, wobei F_0 eine fest vorgegebene Verteilungsfunktion ist.
- (b) H_0 : Die Verteilungsfunktion F gehört zu einer Klasse von Verteilungsfunktionen, die durch m Parameter $\theta_1, \theta_2, \dots, \theta_m$ eindeutig bestimmt ist. Diese Parameter werden aus der Stichprobe geschätzt.

Für die Einteilung in Klassen wird der (i.a. überabzählbare) Wertebereich der Zufallsvariablen X in $r \geq 2$ disjunkte Intervalle zerlegt:

$$A_1 = (-\infty, x_1], A_2 = (x_1, x_2], \dots, A_{r-1} = (x_{r-2}, x_{r-1}], A_r = (x_{r-1}, \infty),$$

mit $x_1 < x_2 < \dots < x_{r-1}$. Die Wahl der Intervalle wird weiter unten erklärt. Analog zum Fall einer Zufallsvariablen mit endlichem Wertebereich werden als beobachtete Häufigkeiten die Klassenhäufigkeiten n_i bestimmt, d.h. die Anzahl der Stichprobenwerte, die in A_i liegen, für $i = 1, \dots, r$.

Unter der Nullhypothese besitzt X die Verteilungsfunktion F_0 , mit deren Hilfe die hypothetischen Klassenwahrscheinlichkeiten $p_i = \mathbb{P}(X \in A_i)$ berechnet werden. Im Fall (b) hängen die p_i von den Parametern $\theta_1, \dots, \theta_m$ ab. (Man könnte also genauer $p_i(\theta_1, \dots, \theta_m)$ schreiben.) Diese unbekannten Parameter müssen zunächst durch Punktschätzer aus den zur Verfügung stehenden Daten geschätzt werden, beispielsweise mit der Momentenmethode, siehe oben. Die Klasseneinteilung ist dabei (wie im endlichen Fall) so vorzunehmen, dass für $r \leq 8$ alle $e_i = np_i \geq 5$ sind, oder dass im Fall $r > 8$ alle $e_i \geq 1$ sind.

Die Testentscheidungsregel lautet wie vorher: lehne H_0 ab, falls $\chi^2 > \chi_{r-m-1;1-\alpha}^2$, sonst nicht. Hierin bezeichnet $\chi_{r-m-1;1-\alpha}^2$ das $(1 - \alpha)$ -Quantil der χ^2 -Verteilung mit $r - m - 1$ Freiheitsgraden.

Beispiel (Test auf Exponentialverteilung). Es wird vermutet, dass die Zeit zwischen der Ankunft zweier Bienen auf einer Kirschblüte exponentialverteilt ist. In einem Experiment traten folgende 50 Zeitdauern auf (in Sekunden):

255	291	299	579	48	1244	550	362	3336	1855
1040	5	2790	1	2503	33	1815	864	647	1366
1902	1562	1626	566	702	476	1618	27	740	1207
270	1228	118	206	194	1675	1558	671	363	27
1673	1153	180	750	374	205	1095	238	2128	483

Zur Testdurchführung sollen die Klasseneinteilung $[0, 150]$, $(150, 300]$, $(300, 600]$, $(600, 1100]$, $(1100, 1600]$, $(1600, \infty)$ und die Irrtumswahrscheinlichkeit $\alpha = 0.05$ benutzt werden. Die Verteilungsfunktion $F(x)$ der Exponentialverteilung lautet $F(x) = 1 - e^{-\lambda x}$ für $x \geq 0$ und $F(x) = 0$ sonst. (Der Parameter λ ist gerade das Reziproke des Erwartungswert der zugehörigen Zufallsvariable.) Aus dem Stichprobenmittel $\bar{x} = 897.96$ erhält man für λ den Punktschätzer $\hat{\lambda} = 1/\bar{x} = 0.00111364$, also $F_0(x) = 1 - e^{-0.00111364 \cdot x}$ für $x \geq 0$. Die Klassenwahrscheinlichkeiten lauten $p_1 = F_0(150)$, $p_2 = F_0(300) - F_0(150)$, $\dots$, $p_5 = F_0(1600) - F_0(1100)$, $p_6 = 1 - F_0(1600)$. Zusammenfassen der relevanten Größen liefert:

Klassen	n_i	$F(x_i)$	p_i	$e_i = 50p_i$
$x \leq 150$	7	0.153839	0.153839	7.6920
$150 < x \leq 300$	9	0.284012	0.130173	6.5087
$300 < x \leq 600$	8	0.487361	0.203349	10.1675
$600 < x \leq 1100$	8	0.706243	0.218882	10.9441
$1100 < x \leq 1600$	7	0.831669	0.125426	6.2713
$1600 < x$	11	1.000000	0.168331	8.4166

Für die Testgröße χ^2 berechnet man also

$$\chi^2 = \frac{(7 - 7.6920)^2}{7.6920} + \frac{(9 - 6.5087)^2}{6.5087} + \dots = 3.148.$$

Der Vergleich mit dem passenden χ^2 -Quantil $\chi_{6-1-1,0.95}^2 = \chi_{4,0.95}^2 = 9.488$ zeigt, dass H_0 nicht verworfen werden kann.