

9. TESTEN VON HYPOTHESEN

9.1. PHILOSOPHIE HINTER DEN TESTS

(9.1.1) Überblick

Ein *statistischer Test* dient dazu, zu entscheiden, ob man aufgrund einer Stichprobe eine bestimmte Annahme über die Grundgesamtheit widerlegen oder bestätigen kann. Diese Annahme nennt man auch die *Nullhypothese* (abgekürzt \mathcal{H}_0), ihr Gegenteil ist die *Alternativhypothese* \mathcal{H}_1 . (9.1.4)
Ein solcher Test hat zwei mögliche Ergebnisse: (9.1.6)

1. Man lehnt \mathcal{H}_0 ab und akzeptiert \mathcal{H}_1 . Bei dieser Entscheidung kann man sich aber irren (*Fehler 1. Art*). Immerhin ist die Wahrscheinlichkeit dafür, eine richtige Hypothese fälschlicherweise abzulehnen, bekannt; sie ist höchstens gleich der so genannten *Irrtumswahrscheinlichkeit* (oder *Signifikanzniveau*) α . (9.1.9)
2. Es besteht aufgrund der Daten kein Anlass, \mathcal{H}_0 zu verwerfen. Dies ist aber noch kein Beweis für die Richtigkeit von \mathcal{H}_0 . (9.1.5)

Eine Ablehnung von \mathcal{H}_0 erfolgt dann, wenn sich aufgrund der Daten ein Resultat ergibt, das bei richtiger Nullhypothese sehr unwahrscheinlich ist. Konkret kommt dies so zustande, dass eine aus den Daten bestimmte Zahl, die *Testgröße*, im so genannten *Verwerfungsbereich* liegt, der in Abhängigkeit von der Irrtumswahrscheinlichkeit α bestimmt wird, wobei oft $\alpha = 5\%$ gesetzt wird; die Werte $\alpha = 1\%$ oder 0.1% kommen ebenfalls vor. (9.1.5)

Je nach Fragestellung kann man einen Test *ein-* oder *zweiseitig* durchführen. (9.1.8)

(9.1.2) Einleitung

Neben dem in Kapitel 8 besprochenen *Schätzen von Parametern* besteht eine zweite Grundaufgabe der beurteilenden Statistik im *Testen von Hypothesen*. Es geht dabei allgemein gesagt (Einzelheiten folgen sogleich) darum, unter Verwendung einer Stichprobe eine gewisse Annahme über die Grundgesamtheit zu widerlegen oder zu bestätigen.

Die hierbei angewandten Methoden nennt man *statistische Tests*. Nun ist es so, dass es sehr viele derartige Testverfahren gibt, von denen man in der Praxis je nach der konkreten Fragestellung und der bereits vorhandenen Information über die Grundgesamtheit ein geeignetes auswählt. Es kann hier aber nicht darum gehen, möglichst viele

statistische Tests zu besprechen. Vielmehr soll in den folgenden Kapiteln das Arbeiten mit solchen Tests anhand von zwei wichtigen Beispielen, dem t -Test und dem χ^2 -Test (die jeweils in mehreren Variationen auftreten), erläutert werden. Danach sollten Sie in der Lage sein, allgemeine Begriffe wie “Nullhypothese” oder “Signifikanzniveau” problemlos zu gebrauchen und auch weitere Testverfahren, wie man sie in der Literatur findet, anzuwenden.

Es ist nun so, dass der Grundgedanke, der hinter diesen statistischen Tests steckt, eigentlich recht einfach ist. In den später zu besprechenden Verfahren wird er aber durch komplizierte Formeln und umfangreiche Tabellen etwas verschleiert. Wir wollen deshalb diese Grundidee in diesem Teil an sehr einfachen und überblickbaren Beispielen erläutern und erst in den folgenden Teilen und Kapitel 10 die für die Praxis wichtigen Verfahren besprechen. Wir beginnen mit einigen allgemeinen Betrachtungen.

(9.1.3) Allgemeines über statistische Tests

Wie bereits erwähnt, dienen statistische Tests zur Untersuchung von *Hypothesen*. Solche Hypothesen sind nichts anderes als Vermutungen, über deren Richtigkeit (oder Falschheit) man Bescheid wissen möchte. Sie können in vielerlei Gestalten auftreten, wie die folgenden, beliebig vermehrbaren Beispiele aufzeigen.

1. Morgen wird es regnen.
2. Glühbirnen der Marke MEGAHELL brennen länger als jene der Marke TURBOGRELL.
3. Dieser Würfel ist “ausgewogen”, d.h., alle sechs Augenzahlen treten mit derselben Wahrscheinlichkeit auf.
4. Das Präparat XY senkt den Blutdruck.

Typisch an diesen Hypothesen ist, dass man sie nie mit Sicherheit bejahen oder verneinen kann. Im Fall der Hypothese Nr. 1 (Wetter) kann man zwar nachträglich feststellen, ob es geregnet hat oder nicht, aber im Voraus lässt sich die Frage nicht mit absoluter Sicherheit beantworten. Um im Fall der Hypothese Nr. 2 (Glühlampen) eine verbindliche Antwort zu erhalten, müsste man die gesamte Produktion untersuchen, was aus praktischen Gründen nicht möglich ist. Man ist also auf Stichproben angewiesen, und im Schluss von der Stichprobe auf die Grundgesamtheit liegt ein Unsicherheitsfaktor.

Man hätte nun gerne irgendwelche *Kriterien*, die es ermöglichen, zu entscheiden, ob man eine bestimmte Hypothese annehmen oder ablehnen soll. In manchen Fällen wird man dies einfach dadurch tun, dass man sagt, aufgrund des “gesunden Menschenverstandes” sei die Antwort offensichtlich. In andern Situationen, vor allem dann, wenn zahlenmäßige Daten vorliegen, muss man rechnerische Methoden anwenden, nämlich die hier zur Diskussion stehenden statistischen Tests.

Es ist nun wichtig, zu realisieren, dass man sich bei einem solchen Entscheid immer auch *irren* kann, und zwar nicht nur, wenn man gefühlsmässig entschieden hat, sondern auch bei Anwendung von statistischen Tests, also von rechnerischen Verfahren (auch wenn sie auf einem Computer laufen). Man kann eine Hypothese ablehnen, obwohl sie richtig war; ebenso ist der umgekehrte Fall möglich. Man sollte sich also hüten zu sagen, diese oder jene Behauptung sei statistisch *bewiesen*.

Das folgende recht schlichte Beispiel versucht zu zeigen, mit welchen Überlegungen eine Hypothese angenommen oder abgelehnt wird und wieso Fehler vorkommen können. Es wird sich später zeigen, dass das verwendete Denkschema auch bei den eigentlichen statistischen Tests dasselbe ist.

Beispiel 9.1.3.A

Ich fahre mit meinem Freund in seinem Auto. Er sagt zu mir: “Schalte bitte Radio DRS 3 ein; Taste 1, glaube ich.” (Mein Freund stellt also die *Hypothese* auf, Taste 1 auf dem Autoradio sei mit DRS 3 belegt.) Ich drücke die besagte Taste — und es erklingt ein Ländler. Natürlich sage ich jetzt: “Du, das ist wohl nicht DRS 3!” (Mit andern Worten, ich lehne die Hypothese ab.) Mein Argument ist natürlich, dass DRS 3 wohl kaum Ländler bringt. Immerhin ist meine Behauptung nicht absolut sicher; durch das Zusammentreffen irgendwelcher mysteriöser Umstände könnte ja auch DRS 3 einmal Ländler spielen. Im Hinblick auf später kann man sagen, dass ich zwar die Hypothese ablehne, dass ich mich dabei aber irren kann.

Hätte mein Kollege aber die Hypothese “Taste 1 = DRS 1” aufgestellt, so hätte der Ländlerklang sicher keinen Anlass dafür gegeben, diese zweite Hypothese abzulehnen, aber — und auch dies ist wichtig — er wäre auch kein Beweis dafür gewesen, dass sie richtig ist, denn es gibt wohl auch andere Sender, die Ländler spielen. ☒

Da man sich also beim Entscheid für oder gegen eine bestimmte Hypothese irren kann, stellt sich die Frage, wie wahrscheinlich denn ein Irrtum sei. Hier zeigt sich nun ein wesentlicher Vorteil der rechnerischen “statistischen Tests” gegenüber einem “gefühlsmässigen Entscheid”. Ein solcher Test erlaubt es nämlich, die Wahrscheinlichkeit eines Fehlentscheids anzugeben. Man kann dann etwa sagen: “Ich lehne die Hypothese ab, und die Wahrscheinlichkeit dafür, dass ich mich irre (d.h., die Hypothese fälschlicherweise ablehne) ist höchstens 5%”.

Wir werden diesen letzten Punkt in (9.1.4.h) an einer Fortsetzung des obigen Beispiels illustrieren. Zunächst wollen wir aber die Überlegungen, die hinter den rechnerischen statistischen Tests stehen, an Beispielen erläutern.

(9.1.4) Ein erstes Beispiel

Beispiel 9.1.4.Aa) Fragestellung und Hypothesen

Ich nehme eine Münze aus meinem Portemonnaie und behaupte, sie sei “ausgewogen”, d.h., beim Werfen hätten “Kopf” und “Zahl” dieselbe Chance. Aufgrund der Symmetrie der Münze leuchtet dies auch ein. Immerhin wäre es aber denkbar, dass sie eine versteckte Unregelmässigkeit aufweisen könnte, die den Ausgang verfälschen würde.

Es stehen sich also zwei Hypothesen gegenüber, die wir mit H_0 und \mathcal{H}_1 abkürzen wollen:

H_0 : Die Münze ist ausgewogen.

\mathcal{H}_1 : Die Münze ist nicht ausgewogen.

b) Prüfung der Hypothesen und Entscheidungsregel

Diese Hypothesen prüfen wir mit einem einfachen “Experiment”: Wir werfen die Münze einige Mal, z.B. — wie wir hier annehmen wollen — 16-mal. Auch bei einer ausgewogenen Münze (also wenn H_0 wahr ist), erwarten wir aber kaum, dass genau achtmal “Kopf” und achtmal “Zahl” herauskommen wird. Ein Ergebnis von siebenmal “Kopf” und neunmal “Zahl” wird uns daher wohl schwerlich an der Richtigkeit von H_0 zweifeln lassen. Sind aber unter den 16 Würfeln nur drei “Köpfe”, so haben wir das Gefühl, mit dieser Münze stimme etwas nicht; wir werden die Hypothese H_0 ablehnen. Denselben Schluss würde man selbstverständlich ziehen, wenn die Anzahl der “Köpfe” noch kleiner wäre. Dies führt auf folgendes Kriterium:

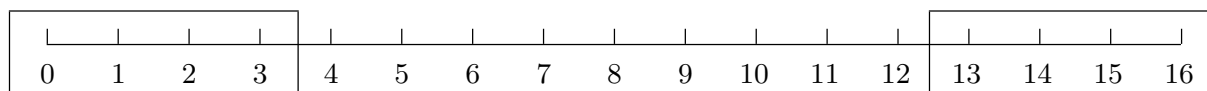
Wenn die Anzahl der “Köpfe” ≤ 3 ist, dann lehnen wir H_0 ab und akzeptieren \mathcal{H}_1

Aus Symmetriegründen sind wir aber gezwungen, dieselbe Schlussfolgerung auch dann zu ziehen, wenn die Anzahl der “Zahlen” ≤ 3 , d.h die Anzahl der “Köpfe” ≥ 13 ist. Zusammenfassend erhalten wir die folgende *Entscheidungsregel*:

Wenn die Anzahl der “Köpfe” ≤ 3 oder ≥ 13 ist, dann lehnen wir die Hypothese H_0 ab und akzeptieren die Hypothese \mathcal{H}_1 .

Unser *Entscheid* lautet dann: Die Münze ist nicht ausgewogen.

Wir stellen das Ganze noch graphisch dar:



H_0 ablehnen

H_0 nicht ablehnen

H_0 ablehnen

Der eingerahmte Bereich heisst *Verwerfungsbereich*, *kritischer Bereich* oder *Ablehnungsbereich*.

c) Wahrscheinlichkeitstheoretische Überlegungen

In der Wahl des Verwerfungsbereichs steckt zunächst noch reine Willkür. Es ist zwar vernünftig, dass er symmetrisch auf beide Enden der Skala verteilt ist, aber es besteht kein einleuchtender Grund dafür, warum die Grenzen bei 3 und 13 und nicht z.B. bei 4 und 12 liegen sollen. Mit Hilfe der Wahrscheinlichkeitsrechnung kann man nun aber die Wahl der Grenzen näher untersuchen.

Dazu formulieren wir unsere beiden Hypothesen H_0 und \mathcal{H}_1 etwas um. Es sei p die Wahrscheinlichkeit dafür, dass bei einem Wurf “Kopf” erscheint. Dann lauten unsere Hypothesen einfach so:

$$\begin{aligned}\mathcal{H}_0 &: p = \frac{1}{2}, \\ \mathcal{H}_1 &: p \neq \frac{1}{2}.\end{aligned}$$

Von jetzt an nennen wir \mathcal{H}_0 auch die *Nullhypothese*, \mathcal{H}_1 die *Alternativhypothese*.

Wir führen jetzt unser “Experiment” durch und werfen die Münze 16-mal. Die Anzahl der auftretenden “Köpfe” ist eine Zufallsgrösse, die wir mit X bezeichnen und die Werte 0, 1, 2, ..., 15, 16 annehmen kann. Da die einzelnen Würfe unabhängig voneinander erfolgen, gehorcht X einer Binomialverteilung (4.2.2) mit dem Parameter $n = 16$. Für die weiteren Untersuchungen nehmen wir an, die Nullhypothese \mathcal{H}_0 sei richtig. Somit gilt für die anderen Parameter p, q der Binomialverteilung, dass $p = q = \frac{1}{2}$ ist. Die übliche Formel für die Wahrscheinlichkeiten (vgl. (4.2.2)) liefert

$$P(X = k) = \binom{16}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{16-k} = \binom{16}{k} \left(\frac{1}{2}\right)^{16}, \quad k = 0, 1, 2, \dots, 16.$$

Mit dieser Formel berechnet man ohne Mühe die folgenden Wahrscheinlichkeiten, wobei natürlich die Symmetrie der Verteilung (vgl. (4.2.2)) ausgenützt wird:

$$\begin{aligned}P(X = 0) &= P(X = 16) = 0.000015 \\ P(X = 1) &= P(X = 15) = 0.000244 \\ P(X = 2) &= P(X = 14) = 0.001831 \\ P(X = 3) &= P(X = 13) = 0.008545 \\ P(X = 4) &= P(X = 12) = 0.027771 \\ P(X = 5) &= P(X = 11) = 0.066650 \\ P(X = 6) &= P(X = 10) = 0.122192 \\ P(X = 7) &= P(X = 9) = 0.174561 \\ P(X = 8) &= 0.196381.\end{aligned}$$

Durch Addition der Wahrscheinlichkeiten können wir nun sofort die Wahrscheinlichkeit dafür ermitteln, dass X im Verwerfungsbereich liegt. Wir bezeichnen dieses Ereignis kurz mit E . Dann gilt

$$P(E) = P(X \in \{0, 1, 2, 3, 13, 14, 15, 16\}) = 0.02127.$$

d) Statistische Entscheidungen und Irrtumswahrscheinlichkeit

Wir fassen unsere bisherigen Überlegungen zusammen:

Wenn die Nullhypothese $\mathcal{H}_0 : p = \frac{1}{2}$ richtig ist, dann ist die Wahrscheinlichkeit dafür, dass X im Verwerfungsbereich liegt, sehr klein, nämlich etwa 2.13%. Das Ereignis E ist also nicht etwa unmöglich, sondern nur sehr wenig wahrscheinlich.

Wir können nun unsere willkürlich getroffene Entscheidungsregel etwas näher unter die Lupe nehmen. Wenn nämlich bei der Durchführung des Versuchs die Anzahl der “Köpfe” im Verwerfungsbereich liegt, dann gibt es nur zwei Möglichkeiten:

1. Die Nullhypothese \mathcal{H}_0 ist falsch.
2. Die Nullhypothese \mathcal{H}_0 ist richtig, aber es ist ein Ereignis eingetreten, das bei richtiger Nullhypothese sehr wenig wahrscheinlich ist. (In unserem Beispiel ein Ereignis mit der Wahrscheinlichkeit 0.02127.)

In der beurteilenden Statistik entscheidet man sich nun in einer solchen Situation stets für die erste Möglichkeit. Man lehnt also die Nullhypothese ab und akzeptiert damit die Alternativhypothese.

Noch etwas anders formuliert: Liegt X im Verwerfungsbereich, so schliesst man, dass \mathcal{H}_0 falsch ist. Man sagt auch “ich lehne \mathcal{H}_0 ab” oder “ich verwerfe \mathcal{H}_0 ” und akzeptiert damit die Alternativhypothese \mathcal{H}_1 .

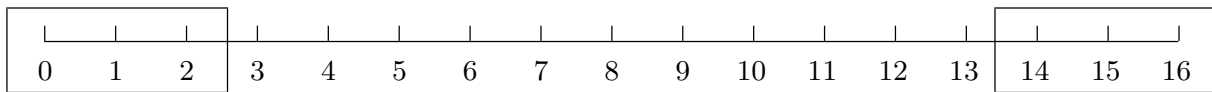
Allerdings muss man sich dabei klar bewusst sein, dass man sich bei diesem Schluss irren kann. Wie wir eben gesehen haben, ist es auch möglich, dass \mathcal{H}_0 richtig ist ($p = \frac{1}{2}$) und dass trotzdem die Anzahl “Köpfe” im Verwerfungsbereich liegt. In einem solchen Fall werden wir \mathcal{H}_0 zu Unrecht verwerfen und somit einen Fehler machen, den wir bewusst in Kauf nehmen, weil die Wahrscheinlichkeit dafür sehr klein ist, nämlich rund 2.13%. Diese Wahrscheinlichkeit heisst *Irrtumswahrscheinlichkeit* oder *Signifikanzniveau* und wird gewöhnlich mit α bezeichnet.

Man kann sich also, wie eben erwähnt, auch bei einem statistischen Test irren. Aussagen wie “Es ist statistisch bewiesen, dass ...” sind somit nicht für bare Münze zu nehmen. Insofern besteht kein Unterschied, ob eine Entscheidung aufgrund eines statistischen Verfahrens oder auf andere Art getroffen wird. Der grosse Vorteil der rechnerischen Tests ist aber, dass man die Wahrscheinlichkeit eines falschen Entscheids zahlenmässig angeben kann, nämlich durch die Irrtumswahrscheinlichkeit α .

e) Irrtumswahrscheinlichkeit und Verwerfungsbereich

Die oben berechnete Irrtumswahrscheinlichkeit $\alpha = 0.02127$ hängt natürlich unmittelbar von der Wahl des Verwerfungsbereichs ab. Ändern wir diesen, ändert sich auch die Grösse α . Dazu zwei Beispiele:

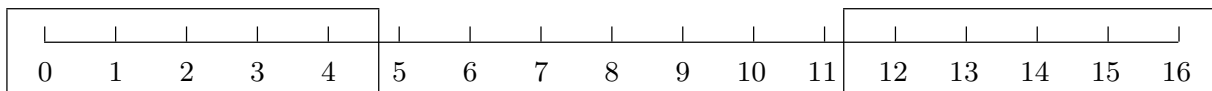
1) Der Verwerfungsbereich sei wie folgt festgelegt:



Hier beträgt die Wahrscheinlichkeit dafür, dass die Anzahl X der “Köpfe” im Verwerfungsbereich liegt (immer unter der Annahme, dass $p = \frac{1}{2}$ ist):

$$P(X \in \{0, 1, 2, 14, 15, 16\}) = 0.00418 \quad (\text{oder } 0.42\%) .$$

2) Für den Verwerfungsbereich



gilt analog

$$P(X \in \{0, 1, 2, 3, 4, 12, 13, 14, 15, 16\}) = 0.076812 \quad (\text{oder } 7.68\%) .$$

f) Zur Festlegung des Verwerfungsbereichs

In den bisherigen Beispielen haben wir immer zuerst den Verwerfungsbereich angegeben und anschliessend die Irrtumswahrscheinlichkeit bestimmt. In der Praxis ist es aber gewöhnlich umgekehrt: Man gibt sich die (maximale) Irrtumswahrscheinlichkeit vor und bestimmt anschliessend den dazu passenden Verwerfungsbereich. In der Wahl dieser Irrtumswahrscheinlichkeit ist man an sich vollkommen frei. Die Person, die den Test durchführt, muss sich von vornherein entscheiden, mit welchem Wert von α sie arbeiten will. Natürlich wird man dieses α , das ja die Wahrscheinlichkeit für eine irrtümliche Ablehnung der Nullhypothese angibt, umso kleiner wählen, je schwerwiegender die Entscheidung ist, die man aufgrund der Beobachtungen fällen muss.

Übliche Irrtumswahrscheinlichkeiten (Signifikanzniveaus) sind

(★)	$\alpha = 0.05$ oder 5% ,
(★★)	$\alpha = 0.01$ oder 1% .
(★★★)	$\alpha = 0.001$ oder 0.1% .

Die Sternchensymbole werden gelegentlich als “Qualitätsangabe” verwendet. In der Biologie ist ein Signifikanzniveau von 5% üblich und sinnvoll; 0.1% sind hier unrealistisch.

Hat man die Wahl von α getroffen, so bestimmt man, wie bereits erwähnt, den zugehörigen Verwerfungsbereich. In unserem Beispiel, oder allgemein immer dann, wenn die Grundgesamtheit durch eine diskrete Zufallsgrösse bestimmt ist, wird es normalerweise nicht möglich sein, den Verwerfungsbereich so zu wählen, dass α genau = 5% (bzw. 1%, 0.1%) wird (bei stetigen Verteilungen geht dies ohne weiteres, wie wir in (9.2.3) noch sehen werden). Man behilft sich hier so, dass man den Verwerfungsbereich möglichst gross wählt, aber so, dass die zugehörige Irrtumswahrscheinlichkeit gerade noch kleiner als das gegebene α ist.

Dazu ein kleines Zahlenbeispiel: Wir haben weiter oben drei Verwerfungsbereiche mit den zugehörigen Wahrscheinlichkeiten angegeben:

$$\begin{aligned} V_1 &= \{0, 1, 2, 14, 15, 16\}, & \alpha &= 0.42\% , \\ V_2 &= \{0, 1, 2, 3, 13, 14, 15, 16\}, & \alpha &= 2.13\% , \\ V_3 &= \{0, 1, 2, 3, 4, 12, 13, 14, 15, 16\}, & \alpha &= 7.68\% . \end{aligned}$$

Mit $\alpha = 5\%$ ist V_2 noch zulässig, nicht aber V_3 , da die Irrtumswahrscheinlichkeit bei V_3 schon 7.68% ist. Wir wählen also V_2 als Verwerfungsbereich. Für $\alpha = 1\%$ würden wir entsprechend V_1 wählen.

g) Anzahl der “Köpfe” nicht im Verwerfungsbereich

Mit einem Signifikanzniveau $\alpha = 5\%$ ist der Verwerfungsbereich gleich

$$V_2 = \{0, 1, 2, 3, 13, 14, 15, 16\} ,$$

wie wir eben gesehen haben. Fallen also in unserem Experiment in 16 Würfeln z.B. 2 oder 13 “Köpfe”, so wird man \mathcal{H}_0 ablehnen, die Münze also als nicht ausgewogen deklarieren, wobei aber immer noch ein (wenn auch wenig wahrscheinlicher) Irrtum passieren kann. Was kann man aber sagen, wenn die Anzahl der “Köpfe” nicht im Verwerfungsbereich liegt, also beispielsweise 7 oder 12 beträgt? Es wäre falsch, zu glauben, in diesem Fall sei \mathcal{H}_0 bewiesen. Vielmehr ist es einfach so, dass wir aufgrund des Testergebnisses \mathcal{H}_0 nicht ablehnen können. Dies bedeutet aber nicht automatisch, dass \mathcal{H}_0 richtig ist. Auch bei einer nicht ausgewogenen Münze ($p \neq \frac{1}{2}$) können schliesslich einmal 7 oder 12 “Köpfe” fallen. Die Berechnung der Wahrscheinlichkeit für ein solches Ereignis muss dann aber für jedes $p \neq \frac{1}{2}$ separat erfolgen; wir gehen nicht näher darauf ein.

h) Ein Vergleich

Wir vergleichen das jetzt sehr ausführlich behandelte Beispiel “Münzenwurf” mit dem in (9.1.3) kurz erwähnten Beispiel “Radio DRS 3”. Wir wollen hier noch zusätzlich annehmen, die Wahrscheinlichkeit dafür, dass DRS 3 in irgendeinem Zeitpunkt Ländl-ermusik spielt, sei bekannt und = 1%. Die Behauptung, Taste 1 des Autoradios sei mit DRS 3 belegt, fassen wir jetzt als Nullhypothese auf. Wenn nun beim Drücken der Taste 1 ein Ländler erklingt, dann gibt es nur zwei Möglichkeiten (wie in d) oben):

1. Die Nullhypothese ist falsch (d.h., der Sender ist nicht DRS 3).
2. Die Nullhypothese ist richtig, aber es ist ein Ereignis eingetreten, das bei richtiger Nullhypothese sehr wenig wahrscheinlich ist (nämlich eines mit einer Wahrscheinlichkeit von 1%).

Genau wie in d) oben entscheiden wir uns in diesem Fall für die Möglichkeit 1 und stellen fest, der Sender sei nicht DRS 3. Mit einer Wahrscheinlichkeit von 1% können wir uns dabei aber irren.

Auch die in g) gemachte Überlegung können wir nachvollziehen: Wenn keine Ländlerklänge ertönen, dann ist es noch lange nicht sicher, dass es sich um DRS 3 handelt; die Tatsache spricht aber auch nicht dagegen.

(9.1.5) Zusammenfassung und Kommentar

Es ist nun an der Zeit, die in bisher angestellten Betrachtungen in allgemeiner Form zusammenzufassen. Wir werden später sehen, dass auch die andern Tests nach demselben Schema aufgebaut sind. Es empfiehlt sich deshalb, diese Zusammenstellung erneut durchzusehen, wenn man in Teil 9.2 bis 9.4 weitere Testverfahren kennen gelernt hat. Nun besprechen wir die einzelnen Schritte.

1. Man stellt die *Nullhypothese* H_0 auf. Das bestmögliche Ergebnis des Tests ist die *Widerlegung* dieser Hypothese. Weiter stellt man die *Alternativhypothese* H_1 auf, die man akzeptiert, wenn H_0 abgelehnt wird.
 2. Man wählt ein Testverfahren.

Kommentar: Im obigen Beispiel besteht der Test im 16-maligen Werfen der Münze.

3. Aufgrund des gewählten Tests bestimmt man mittels der Stichprobe die so genannte *Testgrösse*.

Kommentar: Im allgemeinen Fall ist dies der Wert einer recht komplizierten Zufallsgrösse, den man anhand der Stichprobe mit einer Formel ausrechnet. Wir werden später die Testgrößen t (siehe (9.2.1), (9.3.1)) und χ^2 (siehe (9.4.1)) kennen lernen. Im Beispiel 9.1.4.A ist die Testgrösse einfach der Wert der Zufallsgrösse $X =$ Anzahl "Köpfe".

4. Man wählt ein *Signifikanzniveau* α — meist $\alpha = 5\%$, üblich sind auch noch 1% und 0.1% — und bestimmt aufgrund dessen den *Verwerfungsbereich* (auch *kritischer Bereich* genannt). Dieser wird so festgelegt, dass bei Zutreffen von H_0 die Testgrösse nur mit einer sehr kleinen Wahrscheinlichkeit (nämlich α) im Verwerfungsbereich liegt.

Kommentar: Um diesen Verwerfungsbereich zu bestimmen, muss man die Verteilung der zum Problem gehörenden Zufallsgrösse kennen; dabei ist auch die in der

Nullhypothese getroffene Annahme von Bedeutung. Im Beispiel 9.1.4.A konnten wir diesen Verwerfungsbereich direkt ausrechnen, denn die Testgrösse X folgte einer einfach zu handhabenden Binomialverteilung. Dabei war der eine Parameter n durch den Versuchsaufbau (n Münzenwürfe) und der andere Parameter p durch die Nullhypothese ($p = \frac{1}{2}$) gegeben. Bei komplizierteren Verteilungen ist man für die Bestimmung des Verwerfungsbereichs auf Tabellen angewiesen, wie wir später noch sehen werden.

5. *Entscheidungsregel*: Liegt der gemäss 3. berechnete Wert der Testgrösse im Verwerfungsbereich (gemäss 4.), so lehnen wir \mathcal{H}_0 ab (wir verwerfen die Nullhypothese). Dabei ist es möglich, dass wir \mathcal{H}_0 fälschlicherweise ablehnen, die Wahrscheinlichkeit für dieses Ereignis, die so genannte *Irrtumswahrscheinlichkeit*, ist dabei $\leq \alpha$.

Man sagt in diesem Fall auch, das Ergebnis des Tests sei *signifikant* auf dem 5%-Niveau (oder 1%-Niveau etc.).

Liegt der berechnete Wert der Testgrösse aber *nicht* im Verwerfungsbereich, so können wir \mathcal{H}_0 nicht ablehnen. Dies ist jedoch *kein Beweis* für die Richtigkeit von \mathcal{H}_0 . Vielmehr gibt uns das Testergebnis einfach keinen Anlass, \mathcal{H}_0 zu verwerfen. Eine nicht abgelehnte Nullhypothese darf also nicht unbesehen als richtig akzeptiert werden, sondern kann allenfalls mit der nötigen Vorsicht — z.B. als Arbeitshypothese — weiter verwendet werden.

Es ist wichtig, dass Sie die Grundidee, aufgrund welcher die Entscheidung getroffen wird, stets präsent haben. Kurz zusammengefasst:

Die Nullhypothese wird verworfen, wenn sich aufgrund der Stichprobe ein Resultat ergibt, das bei Gültigkeit dieser Nullhypothese sehr unwahrscheinlich ist.

(9.1.6) Was leistet ein Test und was leistet er nicht?

Wenn Sie die Ausführungen in (9.1.4) und (9.1.5) durchsehen, werden Sie erkennen, dass ein Test (zumindest der oben behandelte, aber die hier zu treffenden Feststellungen gelten für alle Tests) keine Wunder vollbringen kann.

Ein Test kann zwei mögliche Ergebnisse haben:

1. Die Testgrösse liegt im Verwerfungsbereich:
 \mathcal{H}_0 kann verworfen werden.
2. Die Testgrösse liegt nicht im Verwerfungsbereich:
 Es liegt kein Anlass vor, \mathcal{H}_0 zu verwerfen.

Die Aussage 2 ist eigentlich recht schwach, sie darf — wie bereits erwähnt — jedenfalls nicht als Beweis für die Richtigkeit von \mathcal{H}_0 aufgefasst werden.

Die Aussage 1 ist stärker; sie sagt positiv aus, dass man die Nullhypothese ablehnen und damit die Alternativhypothese annehmen kann. Aber auch hier ist die Freude nicht ungetrübt: Es darf nicht vergessen werden, dass die Verwerfung von \mathcal{H}_0 möglicherweise zu Unrecht geschieht. Allerdings hat man die Wahrscheinlichkeit für einen derartigen Irrtum im Griff; sie ist begrenzt durch die gewählte Irrtumswahrscheinlichkeit α .

Aus dem Gesagten ergibt sich ferner, dass man eine Vermutung durch einen Test nicht direkt bestätigen kann. Vielmehr kann man sie bestenfalls indirekt verifizieren, und zwar dann, wenn es möglich ist, ihr Gegenteil als Nullhypothese \mathcal{H}_0 zu formulieren (die ursprüngliche Vermutung ist dann \mathcal{H}_1) und diese mit einem Test abzulehnen.

Es ist aber nicht so, dass man unbesehen jede Behauptung zur Nullhypothese machen kann. Im Beispiel 9.1.4.A des Münzenwurfs etwa ist $\mathcal{H}_0 : p \neq \frac{1}{2}$ nicht zu gebrauchen, denn dann ist der Parameter p der zugrundeliegenden Binomialverteilung gar nicht festgelegt, und wir können die Verteilung von X und damit den Verwerfungsbereich nicht bestimmen.

Schliesslich sei noch auf einige mögliche *Missverständnisse* hingewiesen. Einfachheitshalber arbeiten wir mit $\alpha = 5\%$.

1. Wir nehmen an, die Testgrösse liege im Verwerfungsbereich, wir können also \mathcal{H}_0 ablehnen. Dann ist es unkorrekt, zu sagen:

“ \mathcal{H}_0 ist mit 95% Wahrscheinlichkeit falsch” bzw.

“ \mathcal{H}_1 ist mit 95% Wahrscheinlichkeit richtig”.

Die Hypothese \mathcal{H}_0 (und entsprechend \mathcal{H}_1) ist objektiv gesehen entweder richtig oder falsch (es gibt nichts dazwischen); die Wahrscheinlichkeit bezieht sich nicht auf die Richtigkeit von \mathcal{H}_0 , sondern auf unsere Entscheidung, \mathcal{H}_0 abzulehnen: Hier können wir uns mit 5% Wahrscheinlichkeit irren.

Eine Illustration: Ein nicht geständiger Angeklagter ist entweder schuldig oder unschuldig, nur weiss der Richter nicht, was zutrifft. Die Aussage: “Er ist mit 95% Wahrscheinlichkeit schuldig” bezieht sich also nicht auf den Angeklagten, sondern auf die Meinung des Richters, der zugibt, dass er sich mit 5% Wahrscheinlichkeit irren kann*.

2. Der in 1. angesprochene Sachverhalt zeigt sich noch deutlicher, wenn man beachtet, dass es ohne weiteres vorkommen kann, dass bei einer gegebenen Stichprobe \mathcal{H}_0 bei einer Irrtumswahrscheinlichkeit von 5% verworfen werden kann, nicht aber bei einer solchen von 1%. Die Nullhypothese kann ja nicht das eine Mal falsch und das andere Mal richtig sein. Wieder bezieht sich die Wahrscheinlichkeit auf einen möglichen Fehler bei unserer Entscheidung. Eine Behauptung, die wir mit einer (an sich schon recht geringen) Irrtumswahrscheinlichkeit von 5% zurückweisen können, kann eben möglicherweise bei einer Verkleinerung dieser Wahrscheinlichkeit auf 1% nicht mehr abgelehnt werden.

* Eine verwandte Überlegung finden Sie vor Beispiel 8.4.1.A.

3. Wenn die Testgrösse nicht im Verwerfungsbereich liegt, dann wird man \mathcal{H}_0 beibehalten, da die Daten nicht gegen diese Hypothese sprechen. Hier ist es aber sinnlos, irgendwelche Wahrscheinlichkeiten ins Spiel bringen zu wollen. Insbesondere darf man nicht sagen, \mathcal{H}_0 sei mit 95% Wahrscheinlichkeit richtig. Über die (bedingte) Wahrscheinlichkeit dafür, dass \mathcal{H}_0 richtig ist, falls die Testgrösse nicht im Verwerfungsbereich liegt, hat man keine Information.

(9.1.7) Ein zweites Beispiel

Beispiel 9.1.7.A

Wir besprechen nun einen so genannten *einseitigen* Test, im Gegensatz zum *zweiseitigen* Test. Näheres zum Unterschied finden Sie in (9.1.8).

Wir betrachten dazu wieder unsere Münze, ändern aber die Problemstellung gegenüber Beispiel 9.1.4.A ab. Ich habe nämlich die Vermutung, dass beim Münzenwurf häufiger “Zahl” als “Kopf” erscheint, mit andern Worten, dass $p < \frac{1}{2}$ ist. (Dabei ist p nach wie vor die Wahrscheinlichkeit für “Kopf”.)

Eine erste Frage, die zu klären ist, ist die nach der Nullhypothese. Wir haben in (9.1.5) erkannt, dass man von einem Test als stärkste Aussage erwarten kann, dass \mathcal{H}_0 abgelehnt und H_1 akzeptiert wird. Wir werden deshalb das Gegenteil der Vermutung zur Nullhypothese machen. Diese vielleicht etwas gegen das Gefühl gehende Überlegung ist vor allem bei einseitigen Tests wichtig. Wir setzen also fest:

$$\begin{aligned}\mathcal{H}_0 &: p \geq \frac{1}{2}, \\ H_1 &: p < \frac{1}{2}.\end{aligned}$$

Genau wie in (9.1.3) werfen wir die Münze 16-mal. Damals lautete die Nullhypothese $\mathcal{H}_0 : p = \frac{1}{2}$, und wir lehnten sie ab, wenn die Anzahl X der Köpfe genügend klein bzw. gross war. Bei unserer neuen Fragestellung ist die Sachlage aber anders:

Grosse Werte von X sprechen hier keineswegs gegen die Nullhypothese, die ja behauptet, dass die Wahrscheinlichkeit p für “Kopf” $\geq \frac{1}{2}$ sei. Ein beispielsweise 12-maliges Auftreten von “Kopf” bestätigt ja ganz gewiss die Annahme $p \geq \frac{1}{2}$.

Für den Verwerfungsbereich kommen also hier sicher nur kleine Werte von X in Frage. Um diesen zu bestimmen, übernehmen wir einige der in (9.1.4.c) aufgelisteten Wahrscheinlichkeiten:

$$\begin{aligned}P(X = 0) &= 0.000015 \\ P(X = 1) &= 0.000244 \\ P(X = 2) &= 0.001831 \\ P(X = 3) &= 0.008545 \\ P(X = 4) &= 0.027771 \\ P(X = 5) &= 0.066650.\end{aligned}$$

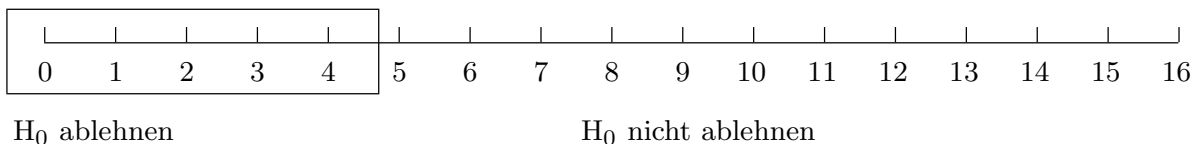
Diese Zahlen basieren allerdings auf der Annahme, dass $p = \frac{1}{2}$ ist, während die neue Nullhypothese lautet: $p \geq \frac{1}{2}$. Nun ist es aber so: Wenn so wenige “Köpfe” auftreten,

dass wir die Hypothese $p = \frac{1}{2}$ zurückweisen können, dann können wir erst recht die Hypothese $p > \frac{1}{2}$ zurückweisen, denn mit $p > \frac{1}{2}$ müssten ja mehr “Köpfe” vorkommen als mit $p = \frac{1}{2}$, und eine kleine Anzahl Köpfe wird daher noch unwahrscheinlicher.

Durch Addition der obigen Wahrscheinlichkeiten erhalten wir (wieder mit $p = \frac{1}{2}$) die folgende Tabelle:

$$\begin{aligned}
 P(X = 0) &= 0.000015 \\
 P(0 \leq X \leq 1) &= 0.000259 \\
 P(0 \leq X \leq 2) &= 0.002090 \\
 P(0 \leq X \leq 3) &= 0.010635 \quad \text{oder} \quad 1.06\% \\
 P(0 \leq X \leq 4) &= 0.038406 \quad \text{oder} \quad 3.84\% \\
 P(0 \leq X \leq 5) &= 0.105056 \quad \text{oder} \quad 10.51\%.
 \end{aligned}$$

Bei einem Signifikanzniveau von 5% muss man also den Verwerfungsbereich wie folgt wählen:



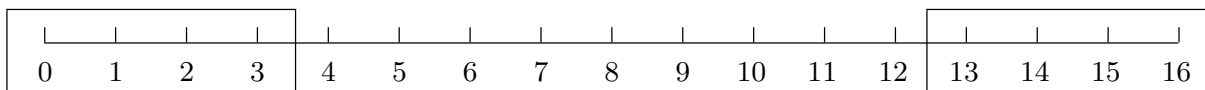
Entsprechend würde man die Verwerfungsbereiche für andere Irrtumswahrscheinlichkeiten bestimmen.

Wir wollen noch zahlenmässig belegen, dass die oben durchgeführte Beschränkung auf den Fall $p = 1/2$ auch den Fall $p > 1/2$ miteinschliesst. Nehmen wir einmal an, es sei $p = 0.55$. Dann rechnet man aus, dass $P(0 \leq X \leq 4) = 1.44\%$ ist. Die Wahrscheinlichkeit dafür, dass X im für $p = 0.5$ bestimmten Verwerfungsbereich $\{0, 1, 2, 3, 4\}$ liegt, ist also wie erwartet kleiner geworden. Dasselbe gilt, wenn wir 0.55 durch eine beliebige Wahrscheinlichkeit p_0 mit $0.5 < p_0 \leq 1$ ersetzen. Wir sehen, dass man in der Tat nicht nur die Hypothese $p = 1/2$, sondern auch die Hypothese $p \geq 1/2$ ablehnen darf, wenn X im angegebenen Verwerfungsbereich liegt.

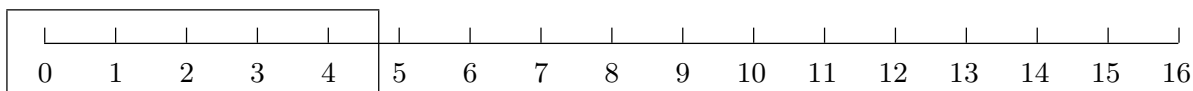
(9.1.8) Einseitige und zweiseitige Tests

Wir vergleichen jetzt die Beispiele 9.1.4.A und 9.1.7.A und geben dazu noch einmal die Hypothesen sowie die Verwerfungsbereiche für $\alpha = 5\%$ an:

9.1.4.A: $\mathcal{H}_0 : p = \frac{1}{2}, \quad \mathcal{H}_1 : p \neq \frac{1}{2}$



9.1.7.A: $\mathcal{H}_0 : p \geq \frac{1}{2}, \quad \mathcal{H}_1 : p < \frac{1}{2}$



Im ersten Fall handelt es sich um eine “zweiseitige Fragestellung”, denn wie man aus der Gestalt des Verwerfungsbereichs erkennt, führen sowohl grosse als auch kleine Werte der Testgrösse zur Ablehnung von H_0 . Man spricht hier von einem *zweiseitigen Test*.

Der zweite Fall aber ist ein Beispiel einer “einseitigen Fragestellung”, denn man interessiert sich nur für Abweichungen nach einer Seite; in diesem Beispiel für Abweichungen nach unten: Es führen nur kleine Werte von X zur Verwerfung von H_0 . Der zugehörige Test wird natürlich *einseitiger Test* genannt. Es ist klar, dass man in manchen Fällen einen einseitigen Test in der umgekehrten Richtung durchführen wird, wobei dann grosse Werte von X zur Ablehnung von H_0 führen.

Wir werden später sehen, dass man auch andere Testverfahren einseitig oder zweiseitig anwenden kann. Die einzelnen Varianten unterscheiden sich dann nur durch die Wahl des Verwerfungsbereichs. Die hier gemachten Angaben übertragen sich sinngemäss.

Der Entscheid, ob ein- oder zweiseitig zu testen sei, ist nicht Sache der Mathematik, sondern die Person, die den Test durchführt, richtet sich nach der Problemstellung. Interessiert sie sich gleichermassen für Abweichungen nach beiden Richtungen, so wird sie zweiseitig, andernfalls einseitig testen. Eine Illustration dazu: Als Kunde bin ich daran interessiert, dass ein Sack Hörnli, den ich kaufe, *mindestens* das angeschriebene Gewicht enthält. Ich würde also einseitig testen. Der Produzent dagegen hat ein Interesse an einem möglichst genauen Gewicht (zuviel bringt ihm Verlust, zuwenig schafft ihm Ärger mit den Konsumentenorganisationen).

Warum verwendet man überhaupt sowohl ein- als auch zweiseitige Tests? Um dies zu erklären, betrachten wir noch einmal die beiden Verwerfungsbereiche, die oben graphisch dargestellt sind. Wir wollen annehmen, unser Versuch habe bei 16 Würfeln gerade vier “Köpfe” ergeben. In diesem Fall könnten wir (immer mit einem Signifikanzniveau von 5%) mit dem zweiseitigen Test die Nullhypothese nicht ablehnen, denn 4 liegt nicht im Verwerfungsbereich. Mit dem einseitigen Test aber können wir H_0 (gerade noch) verwerfen.

Man kann den Unterschied anschaulich so sehen: Beim einseitigen Test konzentriert sich die Irrtumswahrscheinlichkeit von 5% ganz auf das linke Ende der Skala (bzw. in andern Fällen ganz auf das rechte), während beim zweiseitigen Test beide Enden gleichmässig zu berücksichtigen sind, so dass der linke Teil naturgemäss etwas kleiner wird.

Zum Schluss erwähnen wir noch, dass in manchen Büchern auch bei einseitigen Tests die Nullhypothese in der Form $\mathcal{H}_0 : p = 1/2$ (statt $p \geq 1/2$) und analog für andere Tests formuliert wird. Dies wird dadurch gerechtfertigt, dass die Bestimmung des Verwerfungsbereichs tatsächlich unter der Annahme $p = 1/2$ erfolgt (vgl. (9.1.7)). In diesem Fall wird der Unterschied zwischen ein- und zweiseitigem Test allein durch die Formulierung der Alternativhypothese ersichtlich.

(9.1.9) Fehler 1. und 2. Art

Jedem Test liegt eine Nullhypothese \mathcal{H}_0 zugrunde; ihr gegenüber steht die Alternativhypothese \mathcal{H}_1 . Lehnt man \mathcal{H}_0 ab, so akzeptiert man \mathcal{H}_1 . Wie bereits in (9.1.4.d) diskutiert wurde, ist es möglich, dass man aufgrund des Tests \mathcal{H}_0 fälschlicherweise ablehnt. Die Wahrscheinlichkeit dafür ist aber kontrollierbar, sie ist höchstens gleich dem Signifikanzniveau α . Umgekehrt kann es auch vorkommen, dass die Nullhypothese \mathcal{H}_0 beibehalten wird, obwohl sie falsch ist. Man spricht in diesem Zusammenhang von

Fehler 1. Art: Nullhypothese unberechtigterweise abgelehnt,

Fehler 2. Art: Nullhypothese unberechtigterweise beibehalten.

Darstellung in Tabellenform

Ergebnis des Tests \ Wirklichkeit	\mathcal{H}_0 wahr	\mathcal{H}_0 falsch
	\mathcal{H}_0 beibehalten	Entscheid richtig
\mathcal{H}_0 ablehnen	Fehler 1. Art	Entscheid richtig

Illustrationen dazu

- a) In einem Strafprozess hat der Richter zu entscheiden:

\mathcal{H}_0 : XY ist nicht der Täter.

\mathcal{H}_1 : XY ist der Täter.

Ein Fehler 1. Art (\mathcal{H}_0 wahr, aber abgelehnt) bedeutet, dass ein Unschuldiger verurteilt wird.

Ein Fehler 2. Art (\mathcal{H}_0 falsch, aber angenommen) bedeutet, dass ein Schuldiger freigesprochen wird.

Zur Vermeidung eines Justizirrtums ist hier also die Wahrscheinlichkeit für einen Fehler 1. Art möglichst klein zu halten.

- b) Wir betrachten folgende Hypothesen:

\mathcal{H}_0 : Ein neues Medikament hat keine (starken) Nebenwirkungen.

\mathcal{H}_1 : Es hat starke Nebenwirkungen.

Hier muss man versuchen, den Fehler 2. Art sehr klein zu halten, denn ein solcher bedeutet, dass das Medikament trotz möglichen Nebenwirkungen freigegeben würde. Ein Fehler 1. Art dagegen hat nur zur Folge, dass das Medikament nicht in den Handel gelangt.

Wir wissen auch von früher her, dass aufgrund des Prinzips der statistischen Tests die Wahrscheinlichkeit für einen Fehler 1. Art $\leq \alpha$ ist.

Die entsprechende Schranke für den Fehler 2. Art wird mit β bezeichnet. Sie kann aber in der Regel nur unter zusätzlichen Annahmen berechnet werden. Wir betrachten zur Illustration das Beispiel 9.1.4.A mit $\alpha = 5\%$. Wie wir in (9.1.4.f) gesehen haben, ist in diesem Fall der Verwerfungsbereich gleich $\{0, 1, 2, 3, 13, 14, 15, 16\}$. Ein Fehler 2. Art wird begangen, wenn \mathcal{H}_0 falsch (also $p \neq 0.5$) ist, aber X

nicht im Verwerfungsbereich liegt, d.h., wenn $4 \leq X \leq 12$ ist. Nun hängt die Wahrscheinlichkeit für das letztgenannte Ereignis offensichtlich von p ab: Liegt p nahe bei 0.5, so wird sie gross, liegt aber p in der Nähe von 0 oder 1, so wird sie klein sein. Wir werden hier die Wahrscheinlichkeit β nicht näher untersuchen. Immerhin sei bemerkt, dass es bei festem Stichprobenumfang n nicht möglich ist, α und β gleichzeitig so klein zu halten, wie man will. Um α und β beide klein zu machen, muss der Umfang n erhöht werden.

(9.1.10) Ausblick auf die folgenden Teile

In den nächsten Teilen werden einige wichtige Testverfahren besprochen. Die grundlegenden Ideen wurden bereits in (9.1) behandelt und werden uns immer wieder begegnen. Allerdings werden die “technischen” Probleme grösser. Beruhten die Beispiele in diesem Teil einfach auf der Binomialverteilung, so basieren andere Tests auf komplizierteren Verteilungen, wie etwa auf der (stetigen) t -Verteilung.

Die folgenden Teile sind alle nach demselben Schema aufgebaut. Der erste Abschnitt enthält die Beschreibung des Tests, die wie folgt gegliedert ist:

- A. Fragestellung.
- B. Nullhypothese/Alternativhypothese.
- C. Voraussetzungen.
- D. Vorgehen (Testgrösse, Entscheidungsregel).
- E. Bemerkungen.

Es folgen dann jeweils Beispiele, anhand derer auch erläutert wird, wie und warum der Test funktioniert. Für die praktische Durchführung genügt jeweils die Beschreibung, die gewissermassen ein “Kochrezept” ist; zu ihrem tieferen Verständnis ist aber das Studium der Beispiele unerlässlich.

9.2. DER T -TEST FÜR EINE STICHPROBE

(9.2.1) Beschreibung des Tests

A. Fragestellung

Ist der Erwartungswert μ einer Grundgesamtheit gleich oder verschieden von einer gegebenen Zahl μ_0 ?

B. Nullhypothese/Alternativhypothese beim zweiseitigen Test

$$\mathcal{H}_0 : \mu = \mu_0,$$

$$\mathcal{H}_1 : \mu \neq \mu_0.$$

C. Voraussetzungen

1. Die Grundgesamtheit besteht aus Messwerten (im Gegensatz zu Zählwerten), welche — wenigstens im Prinzip — stetig verteilt sind.
2. Die Grundgesamtheit ist normal verteilt. Kleinere Abweichungen von der Normalität werden in der Praxis in Kauf genommen.
3. Der Grundgesamtheit hat man eine Stichprobe vom Umfang n entnommen:

$$x_1, x_2, \dots, x_n.$$

Daraus sind die Grössen \bar{x} und $s_{\bar{x}}$ berechnet worden:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}}.$$

D. Vorgehen beim zweiseitigen Test

1. Man wählt ein Signifikanzniveau α , z.B. $\alpha = 0.05$, und bestimmt aus der Tabelle (6.2.6) (t -Verteilung) den zu α und zum Freiheitsgrad $n - 1$ gehörenden kritischen Wert $t_{\alpha, n-1}$, kurz mit t_α bezeichnet. (Der Freiheitsgrad ist also um 1 kleiner als der Umfang der Stichprobe.)

2. Testgrösse

Anhand der Stichprobe berechnet man die Zahl

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}.$$

3. Entscheidungsregel

- Falls $|t| \geq t_\alpha$ ist, so wird \mathcal{H}_0 verworfen: Die Stichprobe lässt den Schluss zu, dass der Erwartungswert μ signifikant verschieden von μ_0 ist.
- Falls $|t| < t_\alpha$ ist, so besteht aufgrund der Stichprobe kein Anlass, \mathcal{H}_0 zu verwerfen.

E. Bemerkungen

1. Gepaarte Stichproben: Dieser t -Test kann auch bei so genannten “gepaarten” Stichproben (u_i, v_i) , $i = 1, \dots, n$ angewandt werden. Gepaarte Stichproben liegen dann vor, wenn je zwei Messungen am selben Objekt vorgenommen werden. Hier wendet man den t -Test auf die Differenzen $u_i - v_i$ an.
2. Einseitiger Test: Der t -Test kann auch einseitig angewandt werden. Es geht dann um die Frage, ob μ grösser bzw. kleiner als eine gegebene Zahl μ_0 sei. Die Testgrösse t wird genau gleich berechnet. Es ändern sich aber der kritische Wert und die Entscheidungsregel. Das Signifikanzniveau sei wiederum α . Man schliesst wie folgt:

- Mit den Hypothesen

$$\mathcal{H}_0 : \mu \leq \mu_0, \quad \mathcal{H}_1 : \mu > \mu_0$$

wird \mathcal{H}_0 verworfen, wenn $t \geq t_{2\alpha}$ ist.

- ◄ Mit den Hypothesen

$$\mathcal{H}_0 : \mu \geq \mu_0, \quad \mathcal{H}_1 : \mu < \mu_0$$

wird \mathcal{H}_0 verworfen, wenn $t \leq -t_{2\alpha}$ ist.

(9.2.2) Ein erstes Beispiel zur Anwendung des t -TestsBeispiel 9.2.2.A

In diesem Abschnitt geht es darum, anhand eines konkreten Beispiels vorzuführen, wie man das in (9.2.1) beschriebene Verfahren anwendet. In (9.2.3) wird dann am selben Beispiel erklärt, welche Überlegungen eigentlich hinter dem Testverfahren stecken.

Ein Bäcker behauptet: Meine Brötchen wiegen im Durchschnitt genau 70 g. Eine Nachkontrolle von 10 Brötchen ergab folgende Gewichte (in Gramm):

69, 70, 71, 68, 67, 70, 70, 70, 67, 69 .

Der Durchschnitt \bar{x} dieser 10 Gewichte beträgt 69.1 g. Dies widerspricht an sich noch nicht der Behauptung des Bäckers, die sich ja nicht auf die Stichprobe, sondern auf seine gesamte Produktion bezieht. Das von ihm genannte Durchschnittsgewicht von 70 g entspricht vielmehr dem Erwartungswert μ der Grundgesamtheit.

Wir machen die Behauptung des Bäckers zur Nullhypothese:

$$\mathcal{H}_0 : \mu = 70 .$$

Die Zahl μ_0 aus der Beschreibung des Tests ist also hier = 70, während μ wie

eh und je den (unbekannten) Erwartungswert bezeichnet. Ganz generell ist μ_0 in den Anwendungen eine durch die Fragestellung konkret gegebene Zahl.

Die Alternativhypothese ist natürlich

$$H_1 : \mu \neq 70 .$$

Die beiden Hypothesen führen auf einen zweiseitigen Test. Dies ergibt sich auch direkt aus der Problemstellung, da sowohl Gewichtsabweichungen nach unten wie nach oben die Behauptung des Bäckers Lügen strafen.

Zur Berechnung der Testgrösse t benötigen wir noch den Standardfehler $s_{\bar{x}}$. Die in (9.2.1) angegebene Formel liefert sofort $s_{\bar{x}} = 0.4333$. Die Verwendung eines Taschenrechners vereinfacht natürlich die Berechnungen; beachten Sie aber, dass dieser meist die Standardabweichung s liefert (vgl. dazu auch die Bemerkungen zum Thema “Rechner” in (8.2.3) und (8.2.4)). Für den Standardfehler ist s noch durch \sqrt{n} zu dividieren.

Wir fahren nun gemäss Punkt D. (Vorgehen) von (9.2.1) weiter:

1. Wir wählen α wie üblich = 5%. Der Freiheitsgrad $n - 1$ ist hier = $10 - 1 = 9$. Der zur t -Verteilung gehörenden Tabelle (6.2.6) entnehmen wir den kritischen Wert

$$t_{\alpha, n-1} = t_{0.05, 9} = 2.262 ,$$

auch kurz t_α genannt.

2. Mit den erhaltenen Werten für \bar{x} und $s_{\bar{x}}$ berechnen wir die Testgrösse

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - 70}{s_{\bar{x}}} = \frac{69.1 - 70}{0.4333} = -2.077 .$$

3. Entscheid: Es ist $|t| = 2.077 < 2.262 = t_\alpha$. Die Testgrösse t liegt nicht im Verwerfungsbereich. Wir können \mathcal{H}_0 *nicht* zurückweisen:

Die vorliegende Stichprobe spricht *nicht* gegen die Behauptung des Bäckers, das Durchschnittsgewicht seiner Brötchen betrage 70 Gramm.

Damit ist der Test durchgeführt. Die hier beschriebenen Überlegungen und Rechnungen sind etwa die, die man sich beim praktischen Gebrauch macht. Im nächsten Abschnitt (9.2.3) gehen wir dann etwas mehr in die Tiefe.

Zuvor führen wir aber das Beispiel noch etwas weiter. Wählen wir nämlich ein Signifikanzniveau $\alpha = 10\%$, so erhalten wir laut Tabelle den kritischen Wert $t_\alpha = t_{0.1, 9} = 1.833$. Nun entscheiden wir anders, denn weil jetzt $|t| = 2.077 > 1.833 = t_\alpha$ ist, müssen wir H_0 verwerfen. Wir lehnen also die Behauptung des Bäckers als falsch ab.

Die Tatsache, dass wir hier je nach Irrtumswahrscheinlichkeit verschiedene Entscheide fällten, beruht darauf, dass die Irrtumswahrscheinlichkeit sich auf einen möglichen Fehler bei unserm Entscheid bezieht, vgl. hierzu auch (9.1.6). Wenn wir \mathcal{H}_0 zurückweisen, beschuldigen wir den Bäcker der Unlauterkeit; natürlich wollen wir bei einer solchen Aussage eine möglichst grosse Sicherheit haben. Das Zahlenmaterial zeigt

nun, dass wir uns die rufschädigende Aussage nicht leisten dürfen, wenn wir 95% Sicherheit (genauer: eine Irrtumswahrscheinlichkeit von 5%) haben wollen. Sind wir jedoch mit 90% Sicherheit zufrieden, so können wir es riskieren, die besagte Aussage zu machen. \boxtimes

Da der nächste Abschnitt etwas theoretischer wird (unter anderem kommt die t -Verteilung explizit vor), sei zum Schluss noch die anschauliche Bedeutung der Testgrösse $t = (\bar{x} - \mu_0)/s_{\bar{x}}$ beschrieben. Es ist klar, dass uns eine betragsmässig grosse Abweichung des berechneten Durchschnitts \bar{x} vom behaupteten Erwartungswert μ_0 an der Nullhypothese zweifeln lässt. Nun ist aber der Betrag der Abweichung allein noch nicht der richtige Massstab. Wenn nämlich die Grundgesamtheit und damit auch der Durchschnitt stark streut (d.h. eine grosse Standardabweichung hat), dann ist eine bestimmte Differenz $\bar{x} - \mu_0$ weniger verdächtig, als wenn diese Streuung nur gering ist. Aus diesem Grund wird diese Differenz $\bar{x} - \mu_0$ noch durch $s_{\bar{x}}$, die geschätzte Standardabweichung des Durchschnitts (vgl. (8.2.5)), dividiert. So erhält man die Testgrösse t . Wenn diese dem Betrag nach “zu gross” ist (und dies heisst gerade $|t| \geq t_\alpha$), wird \mathcal{H}_0 abgelehnt.

(9.2.3) Etwas zum theoretischen Hintergrund

In diesem Abschnitt nehmen wir nochmals das Beispiel 9.2.2.A auf. Wir wollen sozusagen einen Blick hinter die Kulissen werfen und herausfinden, wie der Test im Einzelnen funktioniert. Eine abgekürzte Version ist soeben am Ende von (9.2.2) gegeben worden.

Wir stellen die Daten aus (9.2.2) erneut zusammen. Wir kennen die Stichprobe (Gewichte von Brötchen)

69, 70, 71, 68, 67, 70, 70, 70, 67, 69 .

Behauptet wird, dass diese Brötchen im Durchschnitt genau 70 g wiegen.

Wir analysieren nun diese Fragestellung etwas genauer. Die *Population* im Sinne von (8.1.3) ist die Menge aller je vom Bäcker hergestellten und herzustellenden Brötchen der zur Diskussion stehenden Sorte; die Anzahl dieser Brötchen ist sehr gross in Bezug auf die ausgewählte Stichprobe. Im Sinne einer Idealisierung denkt man sich diese Menge gewöhnlich sogar unendlich gross. Die *Grundgesamtheit* wird durch die Gewichte der Brötchen beschrieben; diese sind Realisierungen der Zufallsgrösse $X = \text{Gewicht eines Brötchens}$. Der Durchschnitt all dieser Gewichte entspricht dem Erwartungswert μ der Grundgesamtheit (etwas präziser ist $\mu = E(X)$, der Erwartungswert der Zufallsgrösse X). Die Behauptung des Bäckers lautet übersetzt, es sei $\mu = 70$. In der Fragestellung (9.2.1.A) ist also, wie schon in (9.2.2), $\mu_0 = 70$ zu setzen, und die Frage lautet: Ist $\mu = \mu_0 = 70$?

Dies führt wie oben auf die Hypothesen

$$\mathcal{H}_0 : \mu = 70, \quad \mathcal{H}_1 : \mu \neq 70 .$$

Von (9.1.6) wissen wir übrigens, dass als stärkstes Ergebnis allenfalls die Ablehnung von \mathcal{H}_0 resultieren kann.

Wir prüfen noch rasch die in (9.2.1.C) formulierten Voraussetzungen nach: Die erste ist erfüllt, denn das Gewicht ist (im Prinzip) ein stetiges Merkmal, wenn es auch hier durch Runden auf ganze Gramm diskretisiert worden ist. Zur zweiten Voraussetzung ist Folgendes zu sagen: Der Bäcker behauptet ja nicht, dass jedes Brötchen 70 g wiege (dies wäre offensichtlich falsch), sondern nur, dass der Durchschnitt aller Gewichte so gross sei. Da bei der Herstellung der Brötchen viele zufällige Faktoren ins Spiel kommen, darf angenommen werden, dass die Gewichte wenigstens annähernd normal verteilt sind (vgl. (7.2)). Die dritte Voraussetzung bezieht sich auf die Stichprobe; die fraglichen Masszahlen haben wir bereits bestimmt:

$$\bar{x} = 69.1, \quad s_{\bar{x}} = 0.4333 .$$

Das Durchschnittsgewicht der Stichprobe ist also etwas kleiner als die vom Bäcker behauptete Zahl von 70 g. Dies könnte aber Zufall sein: Möglicherweise haben wir aus all den Brötchen von unterschiedlichem Gewicht einfach etwas zuviele leichte erwischt. Genau wie in (9.1.4) besteht nun die Grundidee darin, die Wahrscheinlichkeit dafür zu bestimmen, dass der Durchschnitt der Stichprobe um 0.9 g oder mehr von den behaupteten 70 g abweicht*. Ist diese Wahrscheinlichkeit sehr klein, so wird man die Nullhypothese \mathcal{H}_0 ablehnen, wobei man wie üblich in Kauf nimmt, dass die Ablehnung unberechtigterweise erfolgt.

Nun geht es darum, diesen Ansatz durchzuführen. Wir nehmen dazu an, die Nullhypothese \mathcal{H}_0 sei richtig (genau so haben wir es in (9.1.4) gemacht). Für den Erwartungswert der Grundgesamtheit, also für $\mu = E(X)$ gilt dann

$$\mu = \mu_0 = 70 .$$

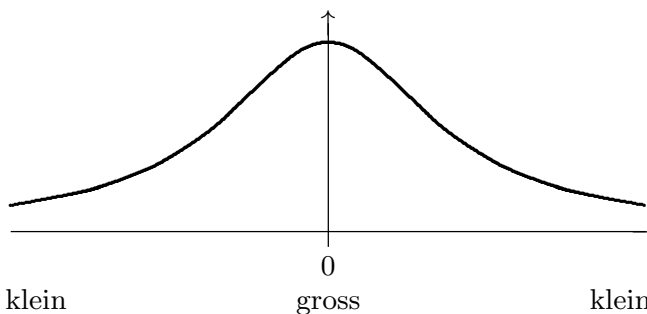
Dieser wird also als bekannt *angenommen*. Nach (8.2.5) ist $\mu = 70$ auch der Mittelwert der Zufallsgrösse \bar{X} . Die Standardabweichung von \bar{X} dagegen ist weiterhin unbekannt; wir müssen sie durch $s_{\bar{x}}$ schätzen (8.2.5). Wie in (8.3) erläutert wurde, geht dadurch die für X und damit für \bar{X} (nach Voraussetzung 9.2.1.C.2) geltende Normalverteilung in eine t -Verteilung über. Dies bedeutet, dass die Testgrösse

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - 70}{s_{\bar{x}}}$$

eine Realisierung einer t -Verteilung mit Freiheitsgrad $n - 1 = 9$ ist.

* Eigentlich bestimmen wir diese Wahrscheinlichkeit gar nicht, sondern modifizieren die Fragestellung etwas. Vgl. die Bemerkung am Ende des Abschnitts (9.2.3).

Wir skizzieren den Graphen dieser Verteilung:

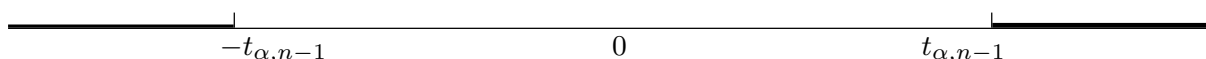


Wie man der Figur entnimmt, ist für extreme Werte von t die Wahrscheinlichkeitsdichte klein. Ergibt sich für die Testgrösse ein solcher Wert, so besteht der begründete Verdacht, die Nullhypothese $\mu = 70$ sei falsch. Liegt aber t nahe beim Nullpunkt, wo die Wahrscheinlichkeitsdichte gross ist, so haben wir keinen Anlass, \mathcal{H}_0 zu verwerfen. (Wir dürfen \mathcal{H}_0 allerdings auch nicht unbesehen annehmen, vgl. (9.1.5).)

Um die beiden Fälle zu trennen, führt man wie in (9.1.4) den *Verwerfungsbereich* ein. Dazu wählt man zuerst ein Signifikanzniveau α und entnimmt der Tabelle (6.2.6) den zu α und zum Freiheitsgrad $n - 1 = 9$ gehörenden kritischen Wert $t_{\alpha, n-1} = t_\alpha$. Wir rufen die in (8.3) erläuterte Bedeutung von $t_{\alpha, n-1}$ in Erinnerung. Diese Zahl ist dadurch charakterisiert, dass

$$P(|T| \geq t_{\alpha, n-1}) = \alpha$$

ist. Die Wahrscheinlichkeit dafür, dass der Wert der Zufallsgrösse T im dick markierten Teil der unten stehenden Figur liegt, ist also $= \alpha$, d.h., sehr klein.



Es ist daher sehr unwahrscheinlich, dass der Wert t der Zufallsgrösse T bei richtiger Nullhypothese \mathcal{H}_0 in diesem Bereich liegt. Trifft dies dennoch zu, so werden wir \mathcal{H}_0 verwerfen, und die Wahrscheinlichkeit dafür, dass wir \mathcal{H}_0 fälschlicherweise ablehnen, ist höchstens gleich α . Der dick markierte Bereich ist also der *Verwerfungsbereich*, wie wir ihn schon in (9.1.4) und (9.1.5) kennen gelernt haben. Die Begründung für die Verwerfung von \mathcal{H}_0 ist genau dieselbe wie in (9.1.4).

Damit haben wir auch die Entscheidungsregel in (9.2.1.D) erklärt: Die Testgrösse t liegt genau dann im Verwerfungsbereich, wenn $|t| \geq t_{\alpha, n-1}$ ist, also gilt

$$|t| \geq t_{\alpha, n-1} \implies \mathcal{H}_0 \text{ ablehnen .}$$

Ist aber $|t| < t_{\alpha, n-1}$, besteht kein Anlass, \mathcal{H}_0 zu verwerfen.

In (9.2.2) haben wir mit konkreten Werten für α gearbeitet und gefunden, dass für den Freiheitsgrad 9

$$t_{0.05} = 2.262 \quad \text{und} \quad t_{0.1} = 1.833$$

ist. Für die Testgrösse t erhielten wir -2.077 . Wir stellen fest, dass $t = -2.077$ für $\alpha = 10\%$ im Verwerfungsbereich liegt (Ablehnung von H_0), nicht aber für $\alpha = 5\%$, und kommen so selbstverständlich zu denselben Schlüssen wie in (9.2.2).

Zusammenfassend lässt sich sagen: Bei Gültigkeit der Nullhypothese folgt die Testgrösse einer t -Verteilung, und unter Verwendung der Tabelle für diese Verteilung können wir sagen, welche Werte von t unwahrscheinlich sind, d.h., den Verwerfungsbereich bilden.

Es sei nochmals betont, dass diese Überlegungen den theoretischen Hintergrund etwas beleuchten sollen. In der Praxis pflegt man — wie in (9.2.2) vorgeführt — einfach nach dem “Kochrezept” (9.2.1) vorzugehen. Ein gewisses Verständnis für den dahintersteckenden Vorgang sollten Sie aber trotz alledem haben.

Der guten Ordnung halber sei schliesslich noch eine kleine Korrektur angebracht: Weiter oben wurde gesagt, die Grundidee des Tests sei es, die Wahrscheinlichkeit dafür zu ermitteln, dass der Durchschnitt der Stichprobe um 0.9 g oder mehr vom durch \mathcal{H}_0 gegebenen Wert 70 g abweiche, dies in Analogie zum Beispiel 9.1.4.A mit der Münze, wo wir direkt gewisse Wahrscheinlichkeiten bestimmt hatten. De facto haben wir hier aber diese Wahrscheinlichkeit nicht bestimmt; vielmehr wird das Problem etwas modifiziert: Man setzt diese Differenz von 0.9 g in die Testgrösse $t = -2.077$ um und prüft nach, ob t im Verwerfungsbereich liegt.

(9.2.4) Ein Beispiel eines einseitigen Tests

Beispiel 9.2.4.A

Wir befassen uns weiterhin mit den Brötchen aus dem Abschnitt (9.2.2). Dort hatten wir gesehen, dass die Nullhypothese “die Brötchen wiegen im Mittel 70 g” auf dem 5%-Niveau nicht zurückgewiesen werden konnte. Etwas anders sieht die Sache aus, wenn wir nachweisen möchten, dass die Brötchen im Durchschnitt zu leicht sind, was man ja aufgrund des Zahlenmaterials vermuten könnte. Formelmässig ausgedrückt möchten wir also zeigen, dass $\mu < 70$ ist. Dies führt auf einen einseitigen Test. Wichtig ist dabei die richtige Wahl von \mathcal{H}_0 und H_1 (vgl. (9.1.5)). Da die stärkste Folgerung aus einem Test die Verwerfung von \mathcal{H}_0 ist, wird man die Vermutung $\mu < 70$ als *Alternativhypothese* wählen:

$$\mathcal{H}_0 : \mu \geq 70, \quad H_1 : \mu < 70 .$$

Wir betrachten genau dieselbe Testgrösse t wie vorhin:

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = -2.077 .$$

Gemäss (9.2.1.E.2) hat sich nun die Entscheidungsregel geändert: Wir vergleichen t mit $t_{2\alpha, n-1}$, kurz auch mit $t_{2\alpha}$ bezeichnet. Der Freiheitsgrad ist immer noch $= 9$. Wenn wir mit $\alpha = 5\%$ arbeiten, dann ist $2\alpha = 10\%$. Wir sehen daher in der Tabelle (6.2.6)

in der oben mit 0.10 beschrifteten Kolonne nach und finden $t_{0.1} = 1.833$ (diese Kolonne ist übrigens in der untersten Zeile der Tabelle mit 0.05 angeschrieben).

Von den beiden in (9.2.1.E.2) aufgeführten Entscheidungsregeln ist aufgrund der aufgestellten Hypothesen die mit ◀ bezeichnete zu verwenden. Da $t \leq -t_{2\alpha}$ ist, werden wir \mathcal{H}_0 verwerfen und damit \mathcal{H}_1 akzeptieren: Wir glauben, dass die Brötchen im Mittel tatsächlich leichter als 70 g sind (bei einer Irrtumswahrscheinlichkeit von 5%).

Der einseitige Test erlaubt es uns also, auf dem 5%-Niveau die Nullhypothese zu verwerfen, was auf demselben Niveau beim zweiseitigen Test nicht der Fall war. \square

Der für die Praxis relevante rechnerische Teil des t -Tests ist damit beschrieben. Wir überlegen uns nun noch, wie die Entscheidungsregel eigentlich zustande kommt.

Die Nullhypothese lautet $\mathcal{H}_0 : \mu \geq 70$. Ein positiver Wert der Testgrösse

$$t = \frac{\bar{x} - 70}{s_{\bar{x}}}$$

bedeutet, dass $\bar{x} > 70$ ist, und dieses Ereignis spricht überhaupt nicht gegen \mathcal{H}_0 , ist also unverdächtig. Wir verwerfen \mathcal{H}_0 nur noch dann, wenn t stark negativ wird. Der Verwerfungsbereich wird somit wie folgt aussehen:

kritischer Wert

Wie ist nun der kritische Wert festzusetzen? Wie üblich arbeiten wir mit $\alpha = 5\%$. Der kritische Wert t_α aus der Tabelle ist dadurch definiert, dass

$$P(|T| \geq t_\alpha) = \alpha$$

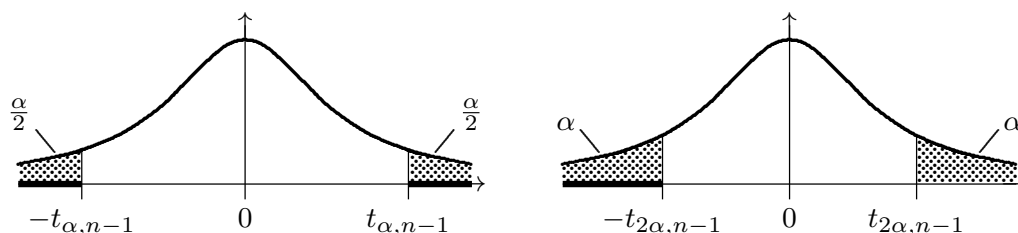
ist, wie wir bereits mehrfach benützt haben. Aus Symmetriegründen ist dann

$$P(T \leq -t_\alpha) = \frac{\alpha}{2} = 0.025 ,$$

$$P(T \geq t_\alpha) = \frac{\alpha}{2} = 0.025 .$$

Die Wahrscheinlichkeit dafür, dass der Wert der Testgrösse in den linken, uns allein noch interessierenden Teil fällt, ist also bloss noch 2.5%. Wir möchten aber, dass diese Wahrscheinlichkeit = 5% wird und sind daher gezwungen, als kritischen Wert den Tabellenwert $t_{0.1}$, also das t_α für $\alpha = 10\%$ zu verwenden.

Allgemein ist beim einseitigen Test mit einem Signifikanzniveau α der Tabellenwert $t_{2\alpha}$ zu benützen. Die folgenden Zeichnungen erläutern den Sachverhalt noch etwas genauer, vgl. auch (5.10.5).



Auch die Entscheidungsregel ist jetzt klar: Positive Werte von t sprechen überhaupt nicht gegen $\mathcal{H}_0 : \mu > 70$. Dagegen führen “stark negative” Werte zur Ablehnung von \mathcal{H}_0 , was die in (9.2.1.E.2) angeführte Regel \blacktriangleleft ergibt.

Selbstverständlich gibt es auch Situationen, wo man einen einseitigen Test in der andern Richtung anwenden wird. Dies geht ganz analog; der Verwerfungsbereich liegt dann rechts des Nullpunkts, und man erhält in (9.2.1.E.2) die Regel \blacktriangleright .

Bemerkungen

1. Beim einseitigen Test haben wir eine Wahl, und zwar zwischen den Hypothesen

$$\blacktriangleright \mathcal{H}_0 : \mu \leq \mu_0, \quad \mathcal{H}_1 : \mu > \mu_0 \quad \text{und} \quad \blacktriangleleft \mathcal{H}_0 : \mu \geq \mu_0, \quad \mathcal{H}_1 : \mu < \mu_0 .$$

Es ist aber unbedingt zu beachten, dass wir beim zweiseitigen Test (9.2.2) diese Wahl nicht haben, die Hypothesen lautet dort zwingend

$$\mathcal{H}_0 : \mu = \mu_0, \quad \mathcal{H}_1 : \mu \neq \mu_0 .$$

Es ist also nicht möglich, als Nullhypothese die Bedingung $\mu \neq \mu_0$ zu wählen und diese allenfalls abzulehnen. Ein ähnlicher Fall wurde in (9.1.6) kurz besprochen.

Der Grund liegt darin, dass die Testgröße t unter der Annahme $\mu = \mu_0$ berechnet werden muss, auch im Fall eines einseitigen Tests.

2. Man kann sich fragen, was geschehen wäre, wenn man in unserm Beispiel \mathcal{H}_0 und \mathcal{H}_1 vertauscht hätte, d.h., wenn man mit

$$\mathcal{H}_0 : \mu \leq 70 \quad \text{und} \quad \mathcal{H}_1 : \mu > 70$$

gearbeitet hätte, also mit der Nullhypothese, die Brötchen seien im Mittel zu leicht. An der Testgröße hätte sich nichts geändert ($t = -2.077$), dagegen hätte mit der Entscheidungsregel \blacktriangleright gearbeitet werden müssen. Man hätte dann einfach herausgefunden, dass \mathcal{H}_0 aufgrund der Stichprobe nicht verworfen werden kann; dies ist aber von allem Anfang an klar, da ja schon $\bar{x} < 70$ ist. Diese Variante des einseitigen Tests bringt also nichts.

3. Obwohl beim einseitigen Test die Nullhypothese besagte, dass $\mu \geq 70$ sei, haben wir die Testgröße t unter der Voraussetzung $\mu = \mu_0 = 70$ berechnet. Wählen wir aber ein größeres μ_0 , so verkleinern wir die Testgröße $t = (\bar{x} - \mu_0)/s_{\bar{x}}$, d.h., diese wird noch stärker negativ. Liegt sie also schon für $\mu_0 = 70$ im Verwerfungsbereich (d.h., weit links), so wird sie dies erst recht für ein größeres μ_0 tun. Liegt t also im Verwerfungsbereich, so können wir nicht nur die Hypothese $\mu = 70$, sondern sogar die Hypothese $\mathcal{H}_0 : \mu \geq 70$ zurückweisen. Ein ähnlicher Fall wurde übrigens in (9.1.7), Ende des Abschnitts, diskutiert.

(9.2.5) Vergleich zweier gepaarter Stichproben

Gepaarte (oder *verbundene*) Stichproben treten dann auf, wenn *am selben Objekt* zwei Messungen derselben Art durchgeführt werden:

- Gewicht vor und nach einer Diät,
- Reaktionszeit vor und nach Alkoholgenuss,
- Blutdruck vor und nach Einnahme eines Medikaments,
- Kraft in der linken und rechten Hand,

usw. Dabei interessiert man sich für die Unterschiede der beiden zusammengehörigen Messungen (Gewichtsabnahme, Verlängerung der Reaktionszeit usw.).

Es seien dazu

$$\begin{aligned} u_1, u_2, \dots, u_n \\ v_1, v_2, \dots, v_n \end{aligned}$$

die beiden Stichproben, wobei u_i und v_i am selben Objekt gemessen wurden. Zu untersuchen sind die Unterschiede

$$x_i = u_i - v_i, \quad i = 1, \dots, n.$$

Dies geschieht dadurch, dass man auf die Grössen x_1, \dots, x_n den t -Test anwendet.

Beim zweiseitigen Test lauten die Hypothesen*

$$\mathcal{H}_0 : \mu = 0, \quad \mathcal{H}_1 : \mu \neq 0.$$

Dabei ist μ der Erwartungswert der Zufallsgrösse $X =$ Differenz der beiden Messwerte. Die Nullhypothese besagt also einfach, es sei keine Veränderung eingetreten; die Alternativhypothese besagt, es sei ein Unterschied (Änderung nach oben *oder* nach unten) festzustellen.

Beim einseitigen Test haben wir die Varianten

$$\begin{aligned} \blacktriangleright \quad \mathcal{H}_0 : \mu \leq 0, \quad \mathcal{H}_1 : \mu > 0 \\ \blacktriangleleft \quad \mathcal{H}_0 : \mu \geq 0, \quad \mathcal{H}_1 : \mu < 0. \end{aligned}$$

Eine Ablehnung von \mathcal{H}_0 im ersten Fall (\blacktriangleright) besagt, dass die Differenzen $u - v$ im Gesamten gesehen positiv sind, d.h., dass beim Übergang von den u -Werten zu den v -Werten eine Abnahme stattgefunden hat. Analog für den zweiten Fall (\blacktriangleleft).

Beispiel 9.2.5.A

Wirkung eines fiebersenkenden Medikaments. In der folgenden Tabelle ist die Körpertemperatur von 10 Patientinnen im Zeitpunkt der Einnahme und drei Stunden nachher angegeben.

* Hier findet die Bezeichnung “Nullhypothese” eine gewisse Begründung.

Nummer des Patientin	Temperatur bei Einnahme	Temperatur nach 3 Std.	Temperaturabnahme
i	u_i	v_i	$x_i = u_i - v_i$
1	39.1	38.7	0.4
2	38.3	38.1	0.2
3	37.6	37.9	-0.3
4	38.0	37.5	0.5
5	40.1	39.2	0.9
6	39.5	39.1	0.4
7	38.7	38.7	0.0
8	37.9	37.5	0.4
9	39.2	38.2	1.0
10	38.0	37.4	0.6

Die Temperatur bei der Einnahme entspricht also den u -Werten, jene nach drei Stunden den v -Werten. Eine *Abnahme* der Temperatur in diesem Zeitraum ergibt daher *positive* x -Werte. Der bei Nr. 3 auftretende negative Wert beschreibt eine Temperaturerhöhung.

Wenn wir nachweisen möchten, dass die Temperatur bei Anwendung des Medikaments *gesenkt* wird, müssen wir *einseitig* testen und zeigen, dass die x -Werte im Mittel positiv sind. Mit derselben Überlegung wie in (9.2.4) wählen wir als Nullhypothese die gegenteilige Aussage (die wir dann im optimalen Fall ablehnen können). Also ist

$$\mathcal{H}_0 : \mu \leq 0, \quad \mathcal{H}_1 : \mu > 0 .$$

Aus den x_i ($i = 1, \dots, 10$) berechnet man sofort

$$\bar{x} = 0.41, \quad s = 0.3872, \quad s_{\bar{x}} = 0.1224, \quad t = \frac{\bar{x} - 0}{s_{\bar{x}}} = \frac{0.41}{0.1224} = 3.350 .$$

(Die Null in der Formel für t kommt von der Nullhypothese her.)

Wir wählen $\alpha = 5\%$ und müssen, da wir einseitig testen, den zu 10% (und dem Freiheitsgrad $10 - 1 = 9$) gehörenden kritischen Wert nachschlagen. Er beträgt 1.833. Da $t = 3.350 > 1.833 = t_{2\alpha}$ ist, können wir nach (9.2.1.E.2) (Variante ►) die Nullhypothese ablehnen und die Alternativhypothese ($\mathcal{H}_1 : \mu > 0$) akzeptieren. Das Mittel hat (bei einer Irrtumswahrscheinlichkeit von 5%) eine fiebersenkende Wirkung.

Wir hätten sogar auf dem 1% -Niveau mit $t_{2\alpha} = t_{0.02} = 2.821 < 3.350 = t$ immer noch Signifikanz erhalten. \square

(9.2.6) Zusammenhang mit dem Konfidenzintervall

Schon die Tatsache, dass sowohl beim t -Test als auch bei der Berechnung des Konfidenzintervalls die t -Verteilung vorkommt, legt den Gedanken nahe, dass ein Zusammenhang zwischen den beiden Themen bestehen wird. Dies ist in der Tat der Fall.

Wir stellen uns dazu die Frage, wann beim zweiseitigen t -Test die Testgrösse t in den Annahmehereich fällt. Dieser ist durch die Beziehung

$$|t| < t_{\alpha, n-1} \quad \text{oder gleichwertig} \quad -t_{\alpha, n-1} < t < t_{\alpha, n-1}$$

gegeben. Setzen wir für t die Formel ein, erhalten wir

$$-t_{\alpha, n-1} < \frac{\bar{x} - \mu_0}{s_{\bar{x}}} < t_{\alpha, n-1} .$$

Äquivalent dazu (Vorzeichenwechsel!) ist

$$-t_{\alpha, n-1} < \frac{\mu_0 - \bar{x}}{s_{\bar{x}}} < t_{\alpha, n-1} .$$

Multiplikation mit $s_{\bar{x}}$ und anschliessende Addition von \bar{x} führt auf

$$\bar{x} - t_{\alpha, n-1} s_{\bar{x}} < \mu_0 < \bar{x} + t_{\alpha, n-1} s_{\bar{x}} .$$

Bis auf etwas andere Bezeichnungen ist dies aber genau die Formel für das Konfidenzintervall (8.4.1). Mit andern Worten: Wir akzeptieren \mathcal{H}_0 genau dann, wenn der in dieser Hypothese vorkommende Wert μ_0 im Konfidenzintervall liegt, das via \bar{x} und $s_{\bar{x}}$ durch die Stichprobe bestimmt wird, wobei für die Vertrauenswahrscheinlichkeit Q gilt: $Q = 1 - \alpha$.

Wir illustrieren den Sachverhalt am in (9.2.2) durchgeführten Test. Mit

$$\bar{x} = 69.1, \quad s_{\bar{x}} = 0.4333, \quad t_{0.05} = 2.262$$

erhalten wir für das Konfidenzintervall ($\alpha = 5\%$, $Q = 95\%$)

$$[\bar{x} - t_{0.05} s_{\bar{x}}, \bar{x} + t_{0.05} s_{\bar{x}}] = [68.12, 70.08] .$$

Der Wert $\mu_0 = 70$ liegt noch in diesem Intervall; wir müssen also $\mathcal{H}_0 : \mu = 70$ akzeptieren (kein Anlass zur Verwerfung), dies in Übereinstimmung mit (9.2.2).

Mit einer Vertrauenswahrscheinlichkeit von 90% (d.h. $\alpha = 10\%$) dagegen verkleinert sich das Intervall wegen $t_{0.1} = 1.833$ auf

$$[68.31, 69.89] .$$

Jetzt liegt $\mu_0 = 70$ nicht mehr in diesem Intervall; in der Tat konnten wir in (9.2.2) die Nullhypothese auf dem Niveau 10% ablehnen.

Diese Übereinstimmung soll aber nicht darüber hinwegtäuschen, dass Vertrauensintervall und t -Test verschiedene Funktionen haben:

- Vertrauensintervall: Angabe einer "Bandbreite" für den *unbekannten* Parameter μ .
- t -Test: Prüfung, ob ein angenommener Erwartungswert μ_0 mit der Stichprobe verträglich ist.

9.3. DER T -TEST FÜR ZWEI UNABHÄNGIGE STICHPROBEN

(9.3.1) Beschreibung des Tests

A. Fragestellung

Sind die Erwartungswerte μ_1, μ_2 zweier Grundgesamtheiten gleich oder verschieden?

B. Nullhypothese/Alternativhypothese beim zweiseitigen Test

$$\mathcal{H}_0 : \mu_1 = \mu_2,$$

$$\mathcal{H}_1 : \mu_1 \neq \mu_2.$$

C. Voraussetzungen

1. Die Grundgesamtheit besteht aus Messwerten (im Gegensatz zu Zählwerten), welche — wenigstens im Prinzip — stetig verteilt sind.
2. Die beiden Grundgesamtheiten sind normal verteilt und haben dieselbe (wenn auch unbekannte) Varianz. Kleinere Abweichungen sowohl der Varianzen als auch von der Normalität werden in der Praxis in Kauf genommen.
3. Der ersten Grundgesamtheit (mit Erwartungswert μ_1) ist eine Stichprobe

$$x_1, x_2, \dots, x_m$$

vom Umfang m , der zweiten (mit Erwartungswert μ_2) eine Stichprobe

$$y_1, y_2, \dots, y_n$$

vom Umfang n entnommen worden.

D. Vorgehen beim zweiseitigen Test

1. Man wählt ein Signifikanzniveau α , z.B. $\alpha = 0.05$, und bestimmt aus der Tabelle (6.2.6) (t -Verteilung) den zu α und zum Freiheitsgrad $m + n - 2$ gehörenden kritischen Wert $t_{\alpha, n-1}$, kurz mit t_α bezeichnet.
2. Testgrösse

Anhand der Stichprobe berechnet man die Zahl

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) \frac{S_{xx} + S_{yy}}{m + n - 2}}}$$

$$\text{mit } \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, S_{xx} = \sum_{i=1}^m (x_i - \bar{x})^2, S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2.$$

(Vgl. (2.2.3.7.a) für die Bezeichnungen S_{xx}, S_{yy} .)

3. Entscheidungsregel

- Falls $|t| \geq t_\alpha$ ist, so wird \mathcal{H}_0 verworfen: Die Stichprobe lässt den Schluss zu, dass μ_1 signifikant verschieden von μ_2 ist.
- Falls $|t| < t_\alpha$ ist, so besteht aufgrund der Stichprobe kein Anlass, \mathcal{H}_0 zu verwerfen.

E. Bemerkungen

Einseitiger Test: Dieser Test kann auch einseitig angewandt werden. Es geht dann um die Frage, ob der eine Erwartungswert grösser als der andere sei. Die Testgrösse t wird genau gleich berechnet. Es ändern sich aber der kritische Wert und die Entscheidungsregel. Das Signifikanzniveau sei wiederum α . Man schliesst wie folgt:

- Mit den Hypothesen

$$\mathcal{H}_0 : \mu_1 \leq \mu_2, \quad \mathcal{H}_1 : \mu_1 > \mu_2$$

wird \mathcal{H}_0 verworfen, wenn $t \geq t_{2\alpha}$ ist.

- ◄ Mit den Hypothesen

$$\mathcal{H}_0 : \mu_1 \geq \mu_2, \quad \mathcal{H}_1 : \mu_1 < \mu_2$$

wird \mathcal{H}_0 verworfen, wenn $t \leq -t_{2\alpha}$ ist.

(9.3.2) Ein erstes Beispiel

Beispiel 9.3.2.A

Wir übernehmen Beispiel 8.1.2.B.

Bei der Züchtung einer neuen Kartoffelsorte fand man in 7 bzw. 6 Versuchsäckern die folgenden Erträge (in kg pro Are):

Alte Sorte ("Alt")	410	420	430	440	450	450	480
Neue Sorte ("Neu")	440	450	455	480	490	505	

Der Züchter möchte natürlich gerne nachweisen, dass die neue Sorte tatsächlich bessere Erträge liefert.

Hier liegen (vgl. auch (8.1.3)) zwei Populationen vor, nämlich alle Ackerstücke, die mit der Sorte "Alt" bzw. der Sorte "Neu" bepflanzt worden sind. Dazu gehören zwei Grundgesamtheiten, nämlich die jeweiligen Erträge pro Hektare.

Mit μ_1 (bzw. μ_2) bezeichnen wir den Erwartungswert der Grundgesamtheit "Alt" (bzw. "Neu"), d.h., die mittleren Erträge pro Hektare der alten bzw. der neuen Sorte,

bezogen auf die gesamten Populationen. Wir möchten zur Freude des Züchters belegen, dass $\mu_1 < \mu_2$ ist. Wir testen deshalb sicher einseitig. Wie üblich beachten wir, dass wir als stärkstes Resultat eines Tests die Nullhypothese ablehnen und die Alternativhypothese akzeptieren können. Wir setzen daher

$$\mathcal{H}_0 : \mu_1 \geq \mu_2, \quad \mathcal{H}_1 : \mu_1 < \mu_2 .$$

(Variante ◀ von (9.3.1.E).)

Die Stichproben stehen in den beiden Zeilen der Tabelle; in der oberen die Werte x_1, x_2, \dots, x_m , in der unteren die Werte y_1, y_2, \dots, y_n . Hier ist also $m = 7, n = 6$. Die beiden Stichproben haben nichts miteinander zu tun, es handelt sich daher um *unabhängige* Stichproben, im Gegensatz zu den in (9.2.5) besprochenen *gepaarten* Stichproben.

Durch Einsetzen in die Formeln berechnet man ohne grosse Mühe

$$\bar{x} = 440, \quad \bar{y} = 470, \quad S_{xx} = 3200, \quad S_{yy} = 3250 .$$

Ein kleiner Tipp zur Benützung eines Rechners: Dieser wird meist keine Taste für S_{xx} haben, wohl aber eine für die Standardabweichung s (bezogen auf die x -Werte). Nach Definition von s ist aber $S_{xx} = (m - 1)s^2$; analog natürlich für die y -Werte.

Nun bestimmen wir die Testgrösse

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) \frac{S_{xx} + S_{yy}}{m + n - 2}}} = \frac{440 - 470}{\sqrt{\left(\frac{1}{7} + \frac{1}{6}\right) \frac{3200 + 3250}{7 + 6 - 2}}} = \dots = -2.2268 .$$

Der Freiheitsgrad ist gleich $7 + 6 - 2 = 11$. Wenn wir das Signifikanzniveau $\alpha = 5\%$ wählen, dann müssen wir, da wir einseitig testen, in der Tabelle (6.2.6) den Wert $t_{2\alpha} = t_{0.1,11}$ nachschlagen. Er ist gleich 1.796, und daher ist $t \leq -t_{2\alpha}$. Die Entscheidungsregel in (9.3.1.E), Variante ◀, erlaubt die Verwerfung von \mathcal{H}_0 . Wir akzeptieren daher die Alternativhypothese $\mathcal{H}_1 : \mu_1 < \mu_2$, die konkret besagt, dass der Ertrag der neuen Sorte tatsächlich grösser als jener der alten Sorte ist.

Beim Signifikanzniveau $\alpha = 1\%$ aber ist der kritische Wert $t_{0.02,11} = 2.718$. Nun ist $t > -2.718$, und wir können \mathcal{H}_0 nicht ablehnen. Die verschiedenen Konklusionen widersprechen sich selbstverständlich nicht. Im ersten Fall ($\alpha = 5\%$) lehnen wir \mathcal{H}_0 ab. Dabei können wir uns aber mit einer Wahrscheinlichkeit von 5% irren. Wenn wir sicherer sein und die Irrtumswahrscheinlichkeit auf 1% senken wollen, dann dürfen wir \mathcal{H}_0 nicht ablehnen; wir haben dann keinen Grund zur Annahme, die neue Sorte liefere höhere Erträge.

Der theoretische Hintergrund, den wir ohne Beweis anführen, ist der, dass die obige, recht komplizierte, Testgrösse t die Realisierung einer Zufallsgrösse ist, welche einer t -Verteilung mit Freiheitsgrad $m + n - 2$ gehorcht, was die Verwendung der Tabelle für die t -Verteilung erklärt.

An diesem Beispiel kann man nochmals die Wahl der Hypothesen beim einseitigen Test illustrieren. Wählen wir die Variante

$$\blacktriangleright \mathcal{H}_0 : \mu_1 \leq \mu_2, \mathcal{H}_1 : \mu_1 > \mu_2 ,$$

so können wir \mathcal{H}_0 sicher nicht zurückweisen (die negative Zahl $t = -2.2268$ ist gewiss nicht $\geq t_{2\alpha} = 1.796$). Dieses Resultat spricht einfach *nicht gegen* $\mathcal{H}_0 : \mu_1 \leq \mu_2$, ist aber auch kein Beweis für die Richtigkeit der Nullhypothese. Mit der Variante \blacktriangleleft haben wir aber gezeigt, dass die Annahme des Gegenteils ($\mu_1 \geq \mu_2$) auf einen Widerspruch (genauer: auf ein sehr unwahrscheinliches Ereignis) führt; als positive Folgerung durften wir weiter oben die dortige \mathcal{H}_1 ($\mu_1 < \mu_2$) annehmen.

(9.3.3) Ein zweites Beispiel

Beispiel 9.3.3.A

Eine Untersuchung befasste sich mit der Brenndauer von Batterien. Wir geben hier nicht die einzelnen Daten, sondern gleich die relevanten Masszahlen an:

- Eine Stichprobe von 60 Batterien der Marke \mathcal{X} ergab einen Durchschnitt von $\bar{x} = 20$ Stunden, ferner war $S_{xx} = 55$.
- Eine entsprechende Stichprobe bestehend aus 40 Batterien der Marke \mathcal{Y} lieferte die Werte $\bar{y} = 19.7$, $S_{yy} = 35$.

Wir fragen uns, ob ein *Unterschied* zwischen den beiden Marken bestehe. Diese Fragestellung ist zweiseitig, deshalb lauten die Hypothesen

$$\mathcal{H}_0 : \mu_1 = \mu_2, \quad \mathcal{H}_1 : \mu_1 \neq \mu_2 .$$

Mit den gegebenen Daten können wir die Testgrösse berechnen:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) \frac{S_{xx} + S_{yy}}{m + n - 2}}} = \frac{20 - 19.7}{\sqrt{\left(\frac{1}{60} + \frac{1}{40}\right) \frac{55 + 35}{60 + 40 - 2}}} = \dots = 1.534 .$$

Mit dem Signifikanzniveau 5% ist der kritische Wert für $\alpha = 0.05$ und für $m + n - 2 = 60 + 40 - 2 = 98$ zu ermitteln. Da unsere Tabelle den Wert für Freiheitsgrad 98 nicht enthält, ersetzen wir ihn bedenkenlos durch jenen für 100 (die Werte für 90 und 100 sind ja fast gleich) und finden daher $t_\alpha = 1.984$. Mit dem berechneten Wert von $t = 1.534$ können wir \mathcal{H}_0 nicht zurückweisen; der Test liefert jedenfalls kein Argument gegen die Behauptung, die Brenndauer der beiden Marken \mathcal{X} und \mathcal{Y} sei gleich.

9.4. DER χ^2 -TEST

(9.4.1) Beschreibung des Tests

A. Fragestellung

Sind beobachtete Häufigkeiten x_i mit theoretisch erwarteten Häufigkeiten t_i verträglich oder nicht verträglich?

B. Nullhypothese/Alternativhypothese

\mathcal{H}_0 : Die Stichprobe stammt aus einer Grundgesamtheit mit einer bestimmten gegebenen Wahrscheinlichkeitsverteilung.

\mathcal{H}_1 : Sie stammt nicht daraus.

C. Voraussetzungen

1. Die beobachteten Grössen sind absolute Häufigkeiten. Es handelt sich um Anzahlen (Zählwerte).
2. Es liegen n Beobachtungen vor (Stichprobe vom Umfang n). Diese werden in k Klassen eingeteilt, und x_i ($i = 1, \dots, k$) sei die Anzahl Werte, welche in die i -te Klasse fallen (beobachtete Häufigkeiten). Ferner sei t_i die Anzahl der Werte, welche gemäss der theoretisch vorgegebenen Verteilung zu dieser i -ten Klasse gehören müssten (erwartete oder theoretische Häufigkeiten).
3. Die erwarteten Häufigkeiten t_i sind alle ≥ 5 .

D. Vorgehen

1. Man wählt ein Signifikanzniveau α und bestimmt aus der Tabelle (6.2.4) den zu α und dem Freiheitsgrad ν gehörenden kritischen Wert $\chi_{\alpha, \nu}^2 = \chi_{\alpha}^2$.
Der Freiheitsgrad ν ist gegeben durch die Anzahl der Klassen, abzüglich der Anzahl der linearen Beziehungen, welche zwischen den x_i bestehen und abzüglich der Anzahl der aus der Stichprobe geschätzten Parameter.

2. Testgrösse

Anhand der Stichprobe und der erwarteten Häufigkeiten berechnet man die Zahl

$$\chi^2 = \sum_{i=1}^k \frac{(x_i - t_i)^2}{t_i} = \sum_{i=1}^k \frac{(\text{beob} - \text{erw})^2}{\text{erw}} .$$

3. Entscheidungsregel

- Falls $\chi^2 \geq \chi_{\alpha, \nu}^2$ ist, so wird \mathcal{H}_0 verworfen: Die Stichprobe stammt nicht aus einer Grundgesamtheit mit der vorgegebenen Verteilung.
- Falls $\chi^2 < \chi_{\alpha, \nu}^2$ ist, so besteht aufgrund der Stichprobe kein Anlass, \mathcal{H}_0 zu verwerfen.

E. Bemerkungen

1. Die theoretischen Häufigkeiten t_i werden mit Hilfe von Wahrscheinlichkeiten berechnet, welche sich ihrerseits aus der gemäss \mathcal{H}_0 erwarteten Verteilung ergeben. Deshalb sind die t_i in der Regel keine ganzen Zahlen. Sie sind aber *nicht* zu runden.
2. Falls die theoretischen Häufigkeiten zu klein sind (vgl. (9.4.1.C.3)), legt man benachbarte Klassen zusammen (d.h., man addiert deren Häufigkeiten), bis alle entstandenen Häufigkeiten ≥ 5 sind. (Es handelt sich hier um eine Faustregel, die in der Literatur auch in etwas abgewandelter Form auftritt.)
3. Der χ^2 -Test hat viele Erscheinungsformen. Deshalb ist diese Beschreibung notwendigerweise recht allgemein gehalten, insbesondere, was die Erläuterung des Freiheitsgrades angeht. Es ist deshalb sinnvoll, sich die verschiedenen Varianten anhand der Beispiele zu merken.
4. Das Symbol χ^2 wird "Chi-Quadrat" ausgesprochen.

(9.4.2) Die χ^2 -Verteilung

Der χ^2 -Test beruht auf folgenden theoretischen Grundlagen:

Die gemäss (9.4.1.D.2) berechnete Testgrösse

$$\chi^2 = \sum_{i=1}^k \frac{(x_i - t_i)^2}{t_i} = \sum_{i=1}^k \frac{(\text{beob} - \text{erw})^2}{\text{erw}}$$

folgt unter der Annahme, dass \mathcal{H}_0 richtig ist, einer Verteilung, die mit wachsendem Stichprobenumfang n gegen die so genannte χ^2 -Verteilung strebt. Die Anwendung dieser Verteilung hat demnach approximativen Charakter, was sich speziell für kleine n auswirkt. Deshalb dürfen die erwarteten Häufigkeiten nicht zu klein sein (9.4.1.C.3).

Wir besprechen nun kurz, was hinter den χ^2 -Verteilungen steckt (wir verwenden den Plural, da es für jeden Freiheitsgrad eine solche Verteilung gibt). Es handelt sich dabei um stetige Verteilungen.

Man geht von der Voraussetzung aus, dass ν Zufallsgrössen X_1, X_2, \dots, X_ν gegeben sind, welche alle der Standard-Normalverteilung $N(0; 1)$ folgen. Wir bilden jetzt die neue Zufallsgrösse*

$$\chi^2 = X_1^2 + \dots + X_\nu^2,$$

deren Verteilung, welche nun eben die χ^2 -Verteilung mit Freiheitsgrad ν genannt wird, bestimmt werden muss. (Ein ähnliches Problem stellte sich im Zusammenhang mit der t -Verteilung in (8.3).) Die zugehörigen Rechnungen können hier nicht durchgeführt werden, immerhin soll das Ergebnis erwähnt sein:

* Das fette χ^2 bezeichnet hier die Zufallsgrösse, das normale χ^2 die Realisierung. Für diese Unterscheidung haben wir sonst Gross- und Kleinbuchstaben verwendet.

Die Dichtefunktion $f(x)$ der χ^2 -Verteilung mit ν Freiheitsgraden ist gegeben durch

$$f(x) = \begin{cases} C_\nu e^{-\frac{x}{2}} x^{\frac{\nu}{2} - 1} & x > 0 \\ 0 & x \leq 0 \end{cases} .$$

Dabei ist C_ν eine Konstante, die hier nicht explizit genannt zu werden braucht. Sie bewirkt, dass

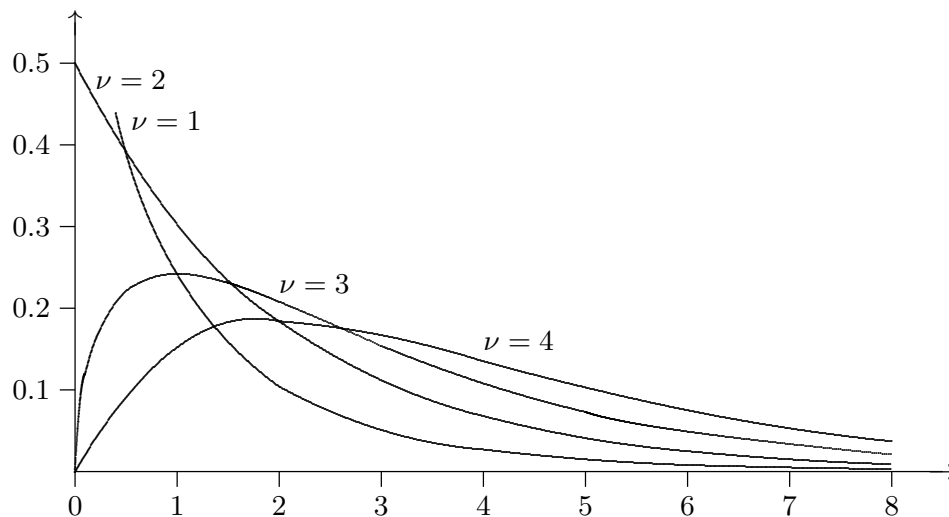
$$\int_{-\infty}^{\infty} f(x) dx = 1$$

ist, wie es bei einer Dichtefunktion sein muss (siehe (4.3.3)).

Infolgedessen ist

$$P(\chi^2 \leq x) = \int_{-\infty}^x f(t) dt .$$

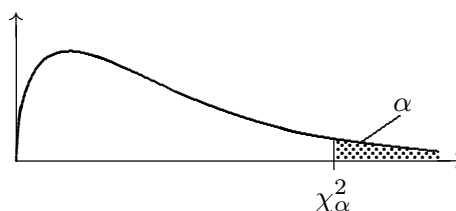
Hier sehen Sie die Graphen der Dichtefunktionen mit $\nu = 1$ bis 4:



Wie schon bei der t -Verteilung interessieren uns nicht so sehr die Werte der Dichtefunktion $f(x)$ oder der Verteilungsfunktion $F(x) = P(\chi^2 \leq x)$, sondern hauptsächlich die *kritischen Werte*. Die Dichtefunktion nimmt für $x < 0$ den Wert 0 an. In diesem Buch ist der Verwerfungsbereich immer “gegen oben”, wie wir gleich nachfolgend illustrieren werden. Für die χ^2 -Verteilung mit Freiheitsgrad ν ist der kritische Wert $\chi_{\alpha, \nu}^2 = \chi_\alpha^2$ definiert durch die Forderung, dass

$$P(\chi^2 \geq \chi_{\alpha, \nu}^2) = \alpha$$

ist. Der Inhalt des hervorgehobenen Flächenstücks ist also $= \alpha$.



Diese kritischen Werte sind in der Tabelle (6.2.4) aufgeführt. In den nächsten Abschnitten werden verschiedene Anwendungen des χ^2 -Tests gegeben.

Die in den Abschnitten (9.4.3) und (9.4.4) vorgestellten Tests heissen auch χ^2 -Anpassungstests, da es um die Frage geht, wie gut sich gegebene Daten an eine theoretische Verteilung anpassen. In (9.4.5) folgt dann noch ein Beispiel eines χ^2 -Unabhängigkeitstests, wo die Unabhängigkeit von Merkmalen untersucht wird.

(9.4.3) Prüfung von Anzahlen auf eine gegebene Verteilung

Beispiel 9.4.3.A

Zur Prüfung eines Würfels werden 60 Würfe ausgeführt. Die beobachteten absoluten Häufigkeiten sind

Augenzahl	1	2	3	4	5	6
Häufigkeit	8	14	8	4	10	16

Theoretisch erwartet man natürlich, dass jede Augenzahl mit der Häufigkeit 10 auftritt. Man spricht hier von einer diskreten Gleichverteilung (vgl. (5.4)). Wir stellen deshalb die folgende Nullhypothese auf:

\mathcal{H}_0 : Die untersuchten Häufigkeiten folgen einer diskreten Gleichverteilung.

Etwas anschaulicher (aber gleichwertig) wäre die Nullhypothese “der Würfel ist unverfälscht”.

Wie in (9.4.1.C) angegeben, bezeichnen wir die beobachteten Häufigkeiten mit x_1, \dots, x_6 , die theoretischen dagegen mit t_1, \dots, t_6 . Die Anzahl der Klassen ist also $k = 6$, während der Umfang der Stichprobe $n = 60$ beträgt. Die beobachteten Häufigkeiten sind gegeben. Für die Berechnung der theoretischen Häufigkeiten gehen wir davon aus, dass \mathcal{H}_0 (Gleichverteilung) zutrifft; es folgt dann sofort, dass alle t_i gleich sind, nämlich = 10.

Zur Berechnung der Testgrösse stellen wir — weil es sich um das erste Beispiel handelt — der Übersichtlichkeit halber folgende Tabelle auf:

i	x_i	t_i	$(x_i - t_i)^2$	$\frac{(x_i - t_i)^2}{t_i}$
1	8	10	4	0.4
2	14	10	16	1.6
3	8	10	4	0.4
4	4	10	36	3.6
5	10	10	0	0.0
6	16	10	36	3.6

Durch Addition der Zahlen in der letzten Kolonne finden wir den Wert der Testgrösse

$$\chi^2 = \sum_{i=1}^6 \frac{(x_i - t_i)^2}{t_i} = \sum_{i=1}^6 \frac{(\text{beob} - \text{erw})^2}{\text{erw}} = 9.6 .$$

Nun ist noch der kritische Wert zu bestimmen. Dazu benötigt man den Freiheitsgrad ν . Wenn es wie hier um die Anpassung an eine Verteilung* geht, dann bestimmt er sich nach der Regel

$$\begin{aligned} \text{Freiheitsgrad} &= \text{Anzahl Klassen} - 1 \\ \nu &= k - 1 \end{aligned}$$

Begründung im Hinblick auf 9.4.1.D.1: Da total 60 Würfe und 6 Klassen vorliegen, sind nur 5 der Häufigkeiten frei wählbar. Die sechste ist dann festgelegt, da die Summe der Häufigkeiten = 60 sein muss: $x_1 + x_2 + \dots + x_6 = 60$. Dies ist eine "lineare Beziehung", da alle Grössen nur in der 1. Potenz vorkommen.

In unserem Fall ist also $\nu = 5$. Mit $\alpha = 5\%$ entnimmt man der Tabelle (6.2.4) den kritischen Wert $\chi_\alpha^2 = 11.070$. Wegen $\chi^2 = 9.6 < 11.07 = \chi_\alpha^2$ liegt keine Signifikanz vor. Wir dürfen die Nullhypothese, die besagt, dass eine Gleichverteilung (bzw. ein unverfälschter Würfel) vorliegt, nicht zurückweisen. Mit andern Worten: Auch bei einem korrekten Würfel ist ein solches Ergebnis in mehr als 5% aller Fälle zu erwarten.

Etwas anders sieht die Sache mit $\alpha = 10\%$ aus. Hier ist $\chi_\alpha^2 = 9.236$, und somit gilt $\chi^2 > \chi_\alpha^2$. Auf diesem Niveau dürfen wir die Nullhypothese ablehnen und behaupten, der Würfel sei verfälscht, wobei aber die Irrtumswahrscheinlichkeit für diese Behauptung 10% beträgt. \boxtimes

Bemerkungen

- 1) Wir können nun auch den anschaulichen Sinn hinter der Testgrösse χ^2 erkennen. Es leuchtet ein, dass die Abweichungen $x_i - t_i$ zwischen den beobachteten und den erwarteten Häufigkeiten zu untersuchen sind. Sicher muss das Vorzeichen weggelassen werden, und wie wir schon früher — etwa in (2.2.3.7.a) — gesehen haben, pflegt man dies durch Quadrieren zu tun. Schliesslich darf aber die Abweichung nicht absolut betrachtet, sondern muss in Beziehung zu den untersuchten Häufigkeiten, also relativ, gesehen werden. Deshalb dividiert man durch t_i und erhält so die Summanden $(x_i - t_i)^2/t_i$.
- 2) Es ist aufgrund der Definition klar, dass χ^2 umso grösser wird, je mehr die beobachteten Häufigkeiten von den erwarteten abweichen. Deshalb sind grosse Werte

* Ohne geschätzte Parameter, vgl. hierzu (9.4.4).

von χ^2 verdächtig, und wie üblich gibt der kritische Wert χ_α^2 gerade die Grenze zwischen “verdächtig” und “unverdächtig” an.

3) Schliesslich sei noch dringend darauf hingewiesen, dass die Nullhypothese

\mathcal{H}_0 : Die Häufigkeiten sind nicht gleich verteilt

nicht verwendet werden kann. Unter dieser Annahme lassen sich nämlich die für die Testgrösse benötigten erwarteten Häufigkeiten gar nicht berechnen, vgl. (9.1.6) für eine ähnliche Situation. Wie in (9.4.1.B) vorgeschrieben, sagt die Nullhypothese stets (also auch in allen folgenden Beispielen) aus, dass eine bestimmte Verteilung vorliegt (und eben nicht, dass sie *nicht* vorliegt).

Beispiel 9.4.3.B

Kreuzt man weiss blühende (Genotyp WW) und rot blühende (Genotyp RR) Erbsen, so erhält man lauter rosa blühende Pflanzen. Kreuzt man dann rosa blühende Pflanzen untereinander, so sollten nach den MENDELSchen Regeln rot, rosa und weiss blühende Erbsen im Verhältnis 1:2:1 auftreten. Ein Versuch ergab die Häufigkeiten 52, 107 und 41. Halten sich diese Abweichungen im Zufallsbereich?

Da es sich um einen Vergleich von Häufigkeiten handelt, können wir den χ^2 -Test anwenden. Die Nullhypothese lautet hier, dass die Stichprobe aus einer Grundgesamtheit von Pflanzen stammt, in der die Verteilung der drei Merkmale (rot, rosa, weiss) durch das Verhältnis 1:2:1 bestimmt ist. Konkreter liesse sich auch sagen, dass hier die MENDELSchen Regeln gelten.

Es liegen total $n = 52 + 107 + 41 = 200$ Beobachtungen vor, aufgeteilt auf drei Klassen ($k = 3$). Gilt die Nullhypothese, so erwartet man 50 rot, 100 rosa und 50 weiss blühende Pflanzen. Man hat also

$$\begin{aligned} x_1 &= 52, & x_2 &= 107, & x_3 &= 41 \\ t_1 &= 50, & t_2 &= 100, & t_3 &= 50 \end{aligned}$$

und berechnet daraus

$$\chi^2 = \frac{(52 - 50)^2}{50} + \frac{(107 - 100)^2}{100} + \frac{(41 - 50)^2}{50} = 2.19 .$$

Nach der im Beispiel 9.4.3.A angegebenen Regel ist der Freiheitsgrad $\nu = 3 - 1 = 2$. Für $\alpha = 0.05$ finden wir $\chi_\alpha^2 = 5.991$. Wegen $\chi^2 < \chi_\alpha^2$ besteht aufgrund der Stichprobe kein Anlass, \mathcal{H}_0 zu verwerfen. Das Versuchsergebnis steht nicht im Widerspruch zu den MENDELSchen Regeln, kann aber auch nicht als Beweis dafür aufgefasst werden. \square

Beispiel 9.4.3.C

Eine statistische Untersuchung von Familien mit 3 Kindern ergab folgende Werte:

Anzahl k der Mädchen	0	1	2	3
Anzahl Familien mit k Mädchen	15	60	95	30

Wir erwarten hier eine Binomialverteilung (siehe (4.2.2)) mit $p = q = \frac{1}{2}$ und $n = 3$. (In diesem Beispiel bezeichnet n wie bei der Binomialverteilung üblich, aber im Gegensatz zu den Bezeichnungen in (9.4.1), die Anzahl der Klassen; n ist also nicht etwa = 200.) Stimmt das wirklich? Wir formulieren die Nullhypothese

\mathcal{H}_0 : Die Stichprobe stammt aus einer Grundgesamtheit, welche binomial verteilt ist, mit den Parametern $n = 3$, $p = q = \frac{1}{2}$.

Mit der bekannten Formel (X bezeichnet die Anzahl der Mädchen)

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, 3$$

berechnet man sofort die Wahrscheinlichkeiten

$$P(X = 0) = \frac{1}{8}, \quad P(X = 1) = \frac{3}{8}, \quad P(X = 2) = \frac{3}{8}, \quad P(X = 3) = \frac{1}{8}.$$

Durch Multiplikation dieser Wahrscheinlichkeiten mit dem Umfang 200 der Stichprobe erhalten wir (wiederum unter Verwendung der Nullhypothese!) die erwarteten Häufigkeiten t_i , die wir zusammen mit den beobachteten Werten x_i angeben:

x_i	15	60	95	30
t_i	25	75	75	25

Daraus berechnet sich die Testgrösse

$$\chi^2 = \frac{(15 - 25)^2}{25} + \frac{(60 - 75)^2}{75} + \frac{(95 - 75)^2}{75} + \frac{(30 - 25)^2}{25} = 13.33.$$

Mit $\nu = 4 - 1 = 3$ und $\alpha = 5\%$ ist $\chi_\alpha^2 = 7.815$. Wegen $\chi^2 > \chi_\alpha^2$ können wir die Nullhypothese ablehnen und (natürlich immer unter Berücksichtigung des Fehlers 1. Art (9.1.9)) sagen, dass die Grundgesamtheit nicht wie behauptet einer Binomialverteilung mit $p = \frac{1}{2}$ folgt. ☒

(9.4.4) Prüfung auf eine gegebene Verteilung mit geschätzten Parametern

Im letzten Beispiel des vorangegangenen Abschnitts haben wir geprüft, ob eine Binomialverteilung mit einem gegebenen Parameter, nämlich $p = \frac{1}{2}$ vorliegt. Eine andere Frage ist die, ob einfach eine Binomialverteilung vorliege, ohne dass über den Parameter p zum vornherein nähere Angaben gemacht werden. Analoge Fragestellungen können natürlich auch bei andern Verteilungen auftreten; wir werden hier Beispiele zur Poisson- und zur Normalverteilung behandeln. Zuerst aber sehen wir uns die Daten aus Beispiel 9.4.3.C von einem andern Blickwinkel aus nochmals an.

Beispiel 9.4.4.A Prüfung auf Binomialverteilung

Wir verwenden dieselben Daten wie in Beispiel 9.4.3.C:

Anzahl k der Mädchen	0	1	2	3
Anzahl Familien mit k Mädchen	15	60	95	30

Wir formulieren nun aber die Nullhypothese anders als vorher:

\mathcal{H}_0 : Die Stichprobe stammt aus einer Grundgesamtheit, welche binomial verteilt ist.

Wir treffen also keine Annahme über den Parameter p (der andere Parameter $n = 3$ ist natürlich gegeben). Um weiter zu kommen, müssen wir nun den Parameter p schätzen. Da in der Tabelle die Anzahl der Mädchen angegeben ist, ist p die Wahrscheinlichkeit einer Mädchengeburt. Der Tabelle entnimmt man, dass von den total $200 \cdot 3 = 600$ Kindern

$$15 \cdot 0 + 60 \cdot 1 + 95 \cdot 2 + 30 \cdot 3 = 340$$

Mädchen waren. Aufgrund der Stichprobe wird man annehmen, der Mädchenanteil in der gesamten Population sei $340/600$ und man wird die Schätzung

$$p = \frac{340}{600} = 0.5667$$

verwenden.

Wie oben verwenden wir nun die Formel

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, 3,$$

diesmal mit $p = 0.5667$, $q = 0.4333$ und finden

$$P(X = 0) = 0.0813, \quad P(X = 1) = 0.3192, \quad P(X = 2) = 0.4175, \quad P(X = 3) = 0.1820.$$

Multiplikation mit $n = 200$ ergibt die theoretischen Häufigkeiten t_i gemäss folgender Tabelle:

x_i	15	60	95	30
t_i	16.26	63.84	83.50	36.40

Die theoretischen Häufigkeiten sind hier keine ganzen Zahlen mehr; man soll sie aber nicht auf- oder abrunden, sondern in dieser Form weiter verwenden, vgl. (9.4.1.E.1). Die Testgrösse χ^2 berechnet sich zu

$$\chi^2 = \frac{(15 - 16.26)^2}{16.26} + \frac{(60 - 63.84)^2}{63.84} + \frac{(95 - 83.50)^2}{83.50} + \frac{(30 - 36.40)^2}{36.40} = 3.0377.$$

Nun zur Bestimmung des Freiheitsgrades. Bei einem χ^2 -Test mit geschätzten Parametern gilt die Regel (vgl. auch (9.4.1.D.1))

$$\text{Freiheitsgrad} = \text{Anzahl Klassen} - \text{Anzahl geschätzter Parameter} - 1.$$

In unserem Falle ist $\nu = 4 - 1 - 1 = 2$, da ein Parameter geschätzt wurde. Mit $\alpha = 5\%$ liefert die Tabelle den kritischen Wert $\chi_\alpha^2 = 5.991$. Wegen $\chi^2 < \chi_\alpha^2$ können wir die Nullhypothese nicht verwerfen. Die Stichprobe spricht nicht dagegen, dass die Grundgesamtheit binomial verteilt ist. \boxtimes

Kommentar

Wodurch unterscheiden sich die beiden Beispiele? Die Nullhypothese von 9.4.3.C (“es liegt eine Binomialverteilung mit $p = \frac{1}{2}$ vor”) besagt mehr als jene von 9.4.4.A (“es liegt eine Binomialverteilung vor”) und ist deshalb leichter zurückzuweisen. In der Tat konnten wir im ersten Fall die Nullhypothese ablehnen, im zweiten nicht.

Beispiel 9.4.4.B Prüfung auf Poisson-Verteilung

Wir benützen die Daten aus Beispiel 8.1.2.D, das hier wiederholt sei:

Im Jahre 1910 zählten RUTHERFORD und GEIGER während 326 Minuten die Zerfälle bei einem radioaktiven Poloniumpräparat, und zwar wurde diese Zeitspanne in Intervalle von 7.5 Sekunden Länge aufgeteilt. In der folgenden Tabelle ist die Anzahl der Intervalle angegeben, in denen 0, 1, 2, ... Zerfälle erfolgten:

Anzahl Zerfälle	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	≥ 15
Anzahl Intervalle	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1	0

Kann man aufgrund dieser Daten annehmen, dass die Zahl der Zerfälle einer Poisson-Verteilung folgt?

Zur Klärung dieser Frage werden wir die Hypothese

$$\mathcal{H}_0 : \text{Die Grundgesamtheit folgt einer Poisson-Verteilung}$$

mit einem χ^2 -Test untersuchen.

Da wir den Parameter λ der Poisson-Verteilung nicht kennen, müssen wir ihn schätzen. Nun wissen wir aber aus (7.3.3), dass λ gerade der Erwartungswert der Verteilung ist, und diesen schätzt man gemäss (8.2.3) mit dem Durchschnitt \bar{x} der Stichprobe. Die 326 Minuten Beobachtungsdauer ergeben total 2608 Zeitintervalle von 7.5 Sekunden Länge. Die Gesamtanzahl der Zerfälle berechnet man (Details seien Ihnen überlassen) zu

$$57 \cdot 0 + 203 \cdot 1 + 383 \cdot 2 + 525 \cdot 3 + \dots + 13 \cdot 1 + 14 \cdot 1 = 10097 .$$

Somit beträgt die durchschnittliche Anzahl der Zerfälle pro Zeitintervall (vgl. auch (8.2.3))

$$\bar{x} = \frac{10097}{2608} = 3.872 ,$$

und daher arbeiten wir mit dem geschätzten Wert $\lambda = 3.872$. Mit Hilfe der üblichen Formel für die Wahrscheinlichkeiten einer Poisson-verteilten Zufallsgrösse

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

berechnet man nun mit etwas Fleiss die Wahrscheinlichkeiten $P(X = k)$ für $k = 0, 1, \dots, 14$. Multipliziert man diese noch mit 2608, der Anzahl der Intervalle, so kommt man auf die theoretischen Häufigkeiten t_i , gemäss folgender Tabelle:

i	0	1	2	3	4	5	6	7	...		
x_i	57	203	383	525	532	408	273	139	...		
t_i	54.29	210.21	406.97	525.26	508.45	393.74	254.10	140.55	...		
...				8	9	10	11	12	13	14	≥ 15
...				45	27	10	4	0	1	1	0
...				68.03	29.27	11.33	3.99	1.29	0.38	0.10	0.04

Hier sind zwei Dinge zu beachten:

1. Die theoretische Häufigkeit für ≥ 15 Zerfälle ist nicht = 0, sondern 0.04. Diese Zahl ergibt sich wenn man die Summe der theoretischen Häufigkeiten t_1, \dots, t_{14} , nämlich 2607.96, auf 2608 ergänzt.
2. In (9.4.1.C.3/E.2) ist die Frage der zu kleinen Werte von t_i erwähnt worden. Wenn gewisse $t_i < 5$ sind, werden benachbarte Klassen zusammengelegt, bis die Häufigkeit gross genug ist. Mit unsern Zahlen bedeutet dies, dass wir die Klassen ab $i = 11$ zusammenziehen müssen*. Wir erhalten dann eine Klasse mit beobachteter Häufigkeit $4 + 0 + 0 + 1 + 1 + 0 = 6$ und erwarteter Häufigkeit $3.99 + 1.29 + 0.38 + 0.10 + 0.04 = 5.80$. Es bleiben dann noch 12 Klassen übrig:

x_i	57	203	383	525	532	408	273	139	45	27	10	6
t_i	54.29	210.21	406.97	525.26	508.45	393.74	254.10	140.55	68.03	29.27	11.33	5.80

Mit etwas Fleiss berechnet man nun die Testgrösse

$$\chi^2 = 12.96 .$$

* Ein anderes Beispiel: Bei fünf Klassen mit den erwarteten Häufigkeiten 3.2, 3.1, 2.4, 2.3 und 2.2 würde man die ersten zwei und die letzten drei zusammenlegen.

Da wir einen Parameter, nämlich λ , geschätzt haben, ist der Freiheitsgrad nach der Regel aus Beispiel 9.4.4.A gleich $12 - 1 - 1 = 10$. Mit $\alpha = 0.05$ finden wir $\chi_\alpha^2 = 18.307$. Wir haben also keinen Anlass, \mathcal{H}_0 zu verwerfen. \square

Beispiel 9.4.4.C Prüfung auf Normalverteilung

In (2.2.2.3) haben wir die Gewichte von Küken angegeben. Kann man aufgrund dieser Daten schliessen, dass die Grundgesamtheit (Gewichte aller zweiwöchigen Küken) normal verteilt ist? Wir können die Hypothese

\mathcal{H}_0 : Die Gewichte folgen einer Normalverteilung

mit einem χ^2 -Test prüfen. Damit die einzelnen Klassen gross genug sind, benützen wir nicht die Klasseneinteilung von Tabelle (2.2.2.3), sondern jene von (2.2.2.4):

Gewicht	beob. Häufigkeit
≤ 90.5	2
90.5 – 95.5	3
95.5 – 100.5	10
100.5 – 105.5	18
105.5 – 110.5	12
110.5 – 115.5	4
> 115.5	1

Um zu kontrollieren, ob eine Normalverteilung vorliegt, müssen wir zuerst die Parameter schätzen. Gemäss (8.2.3) bzw. (8.2.4) wird μ durch \bar{x} und σ durch s geschätzt. Da wir die oben angegebene Klasseneinteilung verwenden, benützen wir konsequenterweise für \bar{x} bzw. s die Formeln für zu Klassen gruppierte Daten. Nach (2.2.3.3.d) ist in diesem Fall $\bar{x} = 103.1$, und nach (2.2.3.7.d) ist $s = 6.267$. Wir runden etwas und verwenden hier die geschätzten Parameter

$$\mu = 103, \sigma = 6.3 .$$

Unter Verwendung der Nullhypothese, die besagt, dass eine Normalverteilung vorliegt, können wir nun berechnen, wieviele Werte theoretisch in den einzelnen Klassen liegen müssten. Die Rechnungen sind etwas umständlich; wir begnügen uns deshalb mit dem Beispiel der Klasse 90.5 – 95.5. Die Wahrscheinlichkeit dafür, dass eine gemäss $N(\mu; \sigma^2)$ verteilte Zufallsgrösse einen Wert im Intervall (90.5, 95.5] annimmt, ist nach (5.10.4) gegeben durch

$$p = P(90.5 < X \leq 95.5) = \Phi_{\mu, \sigma^2}(95.5) - \Phi_{\mu, \sigma^2}(90.5) .$$

Mit den angegebenen Werten von μ und σ berechnen sich die Werte der Verteilungsfunktion Φ_{μ, σ^2} wie folgt (vgl. (5.10.3)):

$$\begin{aligned} \Phi_{\mu, \sigma^2}(95.5) &= \Phi\left(\frac{95.5 - 103}{6.3}\right) = \Phi(-1.19) = 0.1170 , \\ \Phi_{\mu, \sigma^2}(90.5) &= \Phi\left(\frac{90.5 - 103}{6.3}\right) = \Phi(-1.98) = 0.0239 . \end{aligned}$$

Somit ist $p = 0.1170 - 0.0239 = 0.0931$. (Die Werte von Φ können durch Interpolation aus der Tabelle (6.2.3) ermittelt werden.) Da total 50 Küken vorhanden sind, müssten theoretisch in dieser Klasse $50 \cdot p = 50 \cdot 0.0931 = 4.65$ Küken sein.

Ganz entsprechend geht man für die andern Klassen vor. Wir ersparen uns, wie schon erwähnt, die Rechnungen und stellen die Resultate direkt in der folgenden Tabelle dar. Immerhin sei noch darauf hingewiesen, dass für die erste Klasse die Wahrscheinlichkeit $\Phi_{\mu, \sigma^2}(90.5)$ ist; für die letzte Klasse ist $1 - \Phi_{\mu, \sigma^2}(115.5)$. Wir erhalten Folgendes:

Gewicht	beob. Häufigkeit	erw. Häufigkeit
≤ 90.5	2	1.20
90.5 – 95.5	3	4.65
95.5 – 100.5	10	11.38
100.5 – 105.5	18	15.54
105.5 – 110.5	12	11.38
110.5 – 115.5	4	4.65
> 115.5	1	1.20

Die Symmetrie bei den erwarteten Häufigkeiten rührt davon her, dass $\mu = 103$ genau in der Mitte der mittleren Klasse liegt und spielt weiter keine Rolle.

Damit die theoretischen Häufigkeiten ≥ 5 werden, haben wir noch die beiden ersten bzw. die beiden letzten Klassen zusammengefasst, so dass total noch fünf Klassen vorhanden sind.

Mit diesen fünf Klassen berechnen wir die Testgrösse χ^2 :

$$\chi^2 = \sum_{i=1}^5 \frac{(\text{beob} - \text{erw})^2}{\text{erw}} = 0.8376 .$$

Da fünf Klassen und zwei geschätzte Parameter da sind, ist der Freiheitsgrad gemäss der Regel aus Beispiel 9.4.4.A gleich $5 - 1 - 2 = 2$.

Mit $\alpha = 5\%$ bestimmt man $\chi_{\alpha}^2 = 5.991$. Wegen $\chi^2 < \chi_{\alpha}^2$ können wir die Nullhypothese nicht zurückweisen: Die Daten sprechen nicht gegen die Behauptung, die Grundgesamtheit (d.h., das Gewicht von Küken) sei normal verteilt. \boxtimes

(9.4.5) Prüfung einer Vierfeldertafel

In einer *Vierfeldertafel* werden absolute Häufigkeiten nach zwei Merkmalen klassifiziert, wobei jedes Merkmal zwei Ausprägungen hat. Was das genau heisst, soll an zwei Beispielen illustriert werden:

Beispiel 9.4.5.A

In einer grossen Firma werden 500 erwachsene Personen willkürlich herausgegriffen und nach ihren Rauchgewohnheiten befragt. Die Ergebnisse (R : Raucher(in), N :

Nichtraucher(in)) werden in Abhängigkeit vom Geschlecht (M : männlich, W : weiblich) tabelliert (fiktive Zahlen). Ein ähnliche Tabelle haben wir schon in (3.5.2) angetroffen.

	M	W	total
R	250	80	330
N	100	70	170
total	350	150	500

Eine solche Tabelle nennt man auch eine Vierfeldertafel.

Beispiel 9.4.5.B

Es geht darum, eine neue Therapie zu untersuchen. Die Patientengruppe A erhielt die herkömmliche Behandlung, die Gruppe B die neue. Der Erfolg ist in der unten stehenden Vierfeldertabelle vermerkt:

	A	B	total
nicht geheilt	28	22	50
geheilt	155	162	317
total	183	184	367

Diskussion der Beispiele

Im Beispiel 9.4.5.A liegt eine Stichprobe vom Umfang 500 vor. Von den in dieser Stichprobe erfassten Männern rauchen prozentual wesentlich mehr als bei den Frauen (71.4% bei den Männern, 53.3% bei den Frauen). Aufgrund der Stichprobe wird man also annehmen, dass die Rauchgewohnheiten vom Geschlecht abhängig sind, und zwar nicht nur für die Personen in der Stichprobe (wo diese Behauptung unanfechtbar ist), sondern für die ganze Population, d.h., für alle Beschäftigten der Firma. Diese verallgemeinernde Behauptung (Schluss von der Stichprobe auf die Grundgesamtheit) muss nun aber durch einen statistischen Test nachgeprüft werden. Es wird sich weiter unten herausstellen, dass die Behauptung (im Rahmen der Irrtumswahrscheinlichkeit) zulässig ist.

Ähnlich ist die Situation im Beispiel 9.4.5.B. Die Tabelle hinterlässt den Eindruck, die neue Behandlung B sei etwas besser. Darf man diese Schlussfolgerung von der Stichprobe auf die Allgemeinheit übertragen? Der statistische Test gibt die Antwort. Wir werden sehen, dass in diesem Beispiel aufgrund der Daten *nicht* geschlossen werden darf, die neue Therapie sei besser.

Allgemein gesehen geht es darum, nachzuprüfen, ob die in den Zeilen der Vierfeldertafel aufgeführten Merkmale *unabhängig* von jenen in den Spalten sind (oder natürlich umgekehrt die Spaltenmerkmale von den Zeilenmerkmalen, was aufs gleiche herauskommt). Der Begriff der Unabhängigkeit hat ja eine klare Bedeutung in der Wahrscheinlichkeitsrechnung (vgl. (3.5.7)).

Wir gehen deshalb von der folgenden Nullhypothese aus:

\mathcal{H}_0 : Das in den Zeilen beschriebene Merkmal (in den Beispielen das “Rauchverhalten” bzw. der “Heilerfolg”) ist unabhängig von dem in den Spalten beschriebenen (“Geschlecht” bzw. “Heilverfahren”).

Unsere weiteren Überlegungen führen wir am Zahlenmaterial des Beispiels 9.4.5.A vor. Nach (3.5.7) drückt sich die Unabhängigkeit der Ereignisse M und R dadurch aus, dass

$$P(M \cap R) = P(M) \cdot P(R)$$

ist; entsprechend für die drei andern Merkmalskombinationen. Die Wahrscheinlichkeiten $P(M)$ und $P(R)$ sind uns nicht bekannt. Wir können sie aber aufgrund der “Randhäufigkeiten” der Vierfeldertafel schätzen. Von den 500 befragten Personen der Stichprobe sind 350 Männer. Die beste Schätzung, die wir für den Männeranteil der ganzen Belegschaft haben, ist daher 70%. In Formeln

$$P(M) = \frac{350}{500}, \quad P(R) = \frac{330}{500}$$

und entsprechend

$$P(W) = \frac{150}{500}, \quad P(N) = \frac{170}{500}.$$

Setzt man nun die Gültigkeit der Nullhypothese voraus (wie man das bei statistischen Tests immer tut), so sind diese Merkmale unabhängig, und man findet

$$P(M \cap R) = P(M) \cdot P(R) = \frac{350}{500} \cdot \frac{330}{500}.$$

Da die Stichprobe 500 Personen umfasst, ist die *absolute* Häufigkeit (die wir ja beim χ^2 -Test immer verwenden) der rauchenden Männer 500-mal so gross wie diese Wahrscheinlichkeit, also gleich

$$\frac{350 \cdot 330}{500}.$$

Genau gleich verfährt man mit den übrigen drei Fällen. Die Gültigkeit der Nullhypothese (also die Unabhängigkeit der Merkmale) führt auf die folgende Tabelle:

	M	W	total
R	$\frac{350 \cdot 330}{500}$	$\frac{150 \cdot 330}{500}$	330
N	$\frac{350 \cdot 170}{500}$	$\frac{150 \cdot 170}{500}$	170
total	350	150	500

Wir rechnen dies noch aus und erhalten so die erwarteten Häufigkeiten:

	<i>M</i>	<i>W</i>	total
<i>R</i>	231	99	330
<i>N</i>	119	51	170
total	350	150	500

Bemerkungen

- Wie man sieht, erhält man die erwarteten Häufigkeiten dadurch, dass man die jeweiligen “Randhäufigkeiten” multipliziert und durch die Anzahl aller Beobachtungen (den Umfang der Stichprobe) dividiert. Es stört nicht, wenn das Ergebnis keine ganze Zahl ist, vgl. (9.4.1.E.1).
- Die Randhäufigkeiten sind für die beobachteten und die erwarteten Häufigkeiten dieselben. Es genügt daher, nur eine der erwarteten Häufigkeiten gemäss a) zu bestimmen. Die andern lassen sich dann durch Ergänzen auf die Randhäufigkeiten berechnen.
- Die erwarteten Häufigkeiten lassen sich auch durch eine direkte Überlegung, ohne ausdrückliche Erwähnung des Begriffs der Unabhängigkeit, ermitteln: Wenn Rauchverhalten und Geschlecht nichts miteinander zu tun haben, dann müssten sich die total 330 Raucher im Verhältnis 350 (Männer) zu 150 (Frauen) aufteilen, d.h., es müsste

$$\frac{350}{500} \cdot 330$$

rauchende Männer geben; genauso, wie wir es vorhin schon berechnet haben.

Nun vergleichen wir:

Beobachtete Häufigkeiten Erwartete Häufigkeiten

250	80
100	70

231	99
119	51

Mit diesen Häufigkeiten berechnen wir wie üblich χ^2 :

$$\chi^2 = \sum_{i=1}^4 \frac{(\text{beob} - \text{erw})^2}{\text{erw}} = \frac{(250 - 231)^2}{231} + \frac{(80 - 99)^2}{99} + \frac{(100 - 119)^2}{119} + \frac{(70 - 51)^2}{51} = 15.32 .$$

Für die Bestimmung des *Freiheitsgrads* lautet die Regel:

Bei einer Vierfeldertafel ist der Freiheitsgrad = 1

Mit dem üblichen Wert $\alpha = 5\%$ ist $\chi_\alpha^2 = 3.841$. Da χ^2 viel grösser ist, können wir \mathcal{H}_0 zurückweisen: Die Merkmale sind nicht unabhängig. ☒

Eine allgemeine Formel

Die oben dargelegte Methode hat den Vorteil, dass bei der Durchführung jedes Mal klar wird, dass man als Nullhypothese die Unabhängigkeit der Zeilen- bzw. der Spaltenmerkmale gewählt hat. Wer lieber eine abstrakte Formel verwendet, kann dies aber auch haben. Wir schreiben dazu die Vierfeldertafel wie folgt:

a	b	r
c	d	s
t	u	n

wobei gilt

$$r = a + b, \quad s = c + d, \quad t = a + c, \quad u = b + d, \\ n = a + b + c + d = r + s = t + u.$$

Mit diesen Bezeichnungen gilt die folgende Formel:

$$\chi^2 = \frac{(ad - bc)^2 n}{rstu} = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}$$

oder in Worten

$$\chi^2 = \frac{\text{Determinante}^2 \cdot \text{Umfang der Stichprobe}}{\text{Produkt der Randhäufigkeiten}}.$$

(Der Begriff der Determinante wurde in (2.4) im ersten Band kurz erwähnt.)

Mit dieser Formel wollen wir das Beispiel 9.4.5.B (Therapien) durchrechnen. Wir erhalten

$$\chi^2 = \frac{(28 \cdot 162 - 22 \cdot 155)^2 \cdot 367}{50 \cdot 317 \cdot 183 \cdot 184} = 0.872.$$

Bei einem Freiheitsgrad 1 und mit $\alpha = 5\%$ ist $\chi_\alpha^2 = 3.841 > \chi^2$. Wir dürfen daher die Nullhypothese (Unabhängigkeit des Heilerfolgs vom Behandlungsverfahren) nicht zurückweisen. Aufgrund des vorliegenden Datenmaterials darf man nicht schliessen, die neue Behandlung sei besser, obwohl die Gruppe B etwas mehr geheilte Patienten umfasste. \boxtimes

Herleitung der Formel

Zum Beweis der obigen Formel gehen wir genau gleich vor, wie im Beispiel 9.4.5.A. Wir haben die folgende Situation:

Beobachtete Häufigkeiten

a	b
c	d

Erwartete Häufigkeiten

$\frac{rt}{n}$	$\frac{ru}{n}$
$\frac{st}{n}$	$\frac{su}{n}$

Somit ist

$$\chi^2 = \frac{\left(a - \frac{rt}{n}\right)^2}{\frac{rt}{n}} + \frac{\left(b - \frac{ru}{n}\right)^2}{\frac{ru}{n}} + \frac{\left(c - \frac{st}{n}\right)^2}{\frac{st}{n}} + \frac{\left(d - \frac{su}{n}\right)^2}{\frac{su}{n}}.$$

Wir multiplizieren Zähler und Nenner mit n^2 und finden

$$\chi^2 = \frac{(na - rt)^2}{nrt} + \frac{(nb - ru)^2}{nru} + \frac{(nc - st)^2}{nst} + \frac{(nd - su)^2}{nsu}.$$

Im ersten Term setzen wir nun $n = a + b + c + d$, $r = a + b$, $t = a + c$. Rechnet man dies aus, folgt

$$(1) \quad na - rt = (a + b + c + d)a - (a + b)(a + c) = \dots = ad - bc.$$

Ganz analog findet man

$$(2) \quad nb - ru = bc - ad, \quad nc - st = bc - ad, \quad nd - su = ad - bc.$$

Daraus folgt

$$(na - rt)^2 = (nb - ru)^2 = (nc - st)^2 = (nd - su)^2 = (ad - bc)^2.$$

Einsetzen ergibt schliesslich

$$\begin{aligned} \chi^2 &= \frac{(ad - bc)^2}{n} \left(\frac{1}{rt} + \frac{1}{ru} + \frac{1}{st} + \frac{1}{su} \right) = \frac{(ad - bc)^2}{n} \frac{su + st + ru + rt}{rstu} \\ &= \frac{(ad - bc)^2}{n} \frac{(r + s)(t + u)}{rstu} = \frac{(ad - bc)^2}{n} \frac{n \cdot n}{rstu} = \frac{(ad - bc)^2 \cdot n}{rstu}, \end{aligned}$$

womit die Formel bewiesen ist.

Zum Abschluss stellen wir noch fest, dass man den Formeln (1), (2) entnehmen kann, dass die vier Differenzen "beobachtet minus erwartet" bis auf das Vorzeichen alle gleich sind. Im Beispiel 9.2.5.A etwa war der Betrag dieser Differenzen stets = 19.

(9.∞) Aufgaben

9-1 Anlässlich einer Lotterie mit fortlaufend (1,2,...) nummerierten Losen kaufe ich 5 der gut gemischten Lose. Sie tragen die Nummern 777, 1291, 1600, 1492 und 800. Die Losverkäuferin behauptet, es seien mindestens 3000 Lose in Verkauf. Es irritiert mich deshalb etwas, dass ich lauter Nummern ≤ 1600 erwische habe.

Testen Sie die Nullhypothese

$$H_0 : \text{Anzahl der Lose} \geq 3000$$

gegen die Alternativhypothese

$$H_1 : \text{Anzahl der Lose} < 3000.$$

- 9–2 Neun Kolleg(inn)en sind mit ihrem Kleinbus ins Ausland gefahren und haben eingekauft. Fünf sind ehrlich, vier schmuggeln. Wie es das Schicksal so will, kommt es zu einer Kontrolle. Der Zollbeamte wählt drei Personen aus: Alle drei haben geschmuggelt.
- Testen Sie die Nullhypothese H_0 : “Die Auswahl erfolgte zufällig” gegen die Alternativhypothese H_1 : “Der Beamte hat Talent (und verdient deshalb, befördert zu werden)”.
 - Geben Sie die konkrete Bedeutung von “Fehler 1. Art” und “Fehler 2. Art” an.
- 9–3 Beim fünfmaligen Werfen einer Münze ist jedes Mal dieselbe Seite erschienen. Gefühlsmässig würde man vielleicht vermuten, mit der Münze sei etwas nicht in Ordnung. Zeigen Sie aber, dass man die Nullhypothese “die Münze ist ausgewogen” gegenüber der Alternativhypothese “sie ist nicht ausgewogen” auf dem 5%-Niveau *nicht* verwerfen darf. (Es braucht also recht viel, bis H_0 verworfen werden kann!)
- 9–4 Bei einem ehrlichen Würfel ist die Wahrscheinlichkeit für eine Sechs = $\frac{1}{6}$. Ich habe den Verdacht, dass beim Würfel meines Partners die Sechs mit einer grösseren Wahrscheinlichkeit erscheint. Ein Versuch ergibt bei 6 Würfeln 3 Sechsen. Kann ich die Hypothese $H_0 : p \leq \frac{1}{6}$ zurückweisen
- auf dem 5%-Niveau,
 - auf dem 10%-Niveau?
- Geben Sie ferner die konkrete Bedeutung von “Fehler 1. Art” und “Fehler 2. Art” an.
- 9–5 Mein Kollege behauptet, hellsehen zu können. Ich will diese Behauptung testen und lege ihm 12 französische Spielkarten verdeckt hin. Bei jeder Karte muss er sagen, ob sie rot oder schwarz ist. Die Nullhypothese H_0 lautet: Mein Kollege kann *nicht* hellsehen. Wieviele richtige Antworten muss er mindestens geben, damit ich H_0 ablehne (und damit an seine Fähigkeit glaube), wenn ich mich mit höchstens a) 5%, b) 10% Wahrscheinlichkeit irren will?
- Geben Sie ferner die konkrete Bedeutung von “Fehler 1. Art” und “Fehler 2. Art” an.
- 9–6 Eine Maschine produziert Nägel, deren Länge als stetige, normal verteilte Zufallsgrösse X aufgefasst werden kann. Der Erwartungswert μ hängt von der Einstellung der Maschine ab, die Standardabweichung $\sigma = 0.5$ (mm) sei eine feste Grösse. Der Sollwert der Nägel beträgt 50 mm. Zur Kontrolle, ob die Maschine richtig eingestellt ist, werden 100 Nägel zufällig ausgewählt und gemessen. Der Durchschnitt dieser Werte ist die Realisierung der Zufallsgrösse \bar{X} . Zu testen ist also $H_0 : \mu = 50$ gegen $H_1 : \mu \neq 50$. Für welche Werte von \bar{X} muss man H_0 bei einem Signifikanzniveau von 5% verwerfen?
- 9–7 Ein Hersteller liefert an Kioske Schachteln mit je 100 Wundertüten. Es gibt zwei Sortimente. In Sortiment A enthält eine Tüte mit 10% Wahrscheinlichkeit einen Gutschein für den Bezug einer weiteren Tüte; im Sortiment B dagegen nur mit 2% Wahrscheinlichkeit. Nun sind die Schachteln versehentlich nicht angeschrieben worden. Die Kioskinhaberin öffnet daraufhin 10 Tüten. Findet sie keinen Gutschein, nimmt sie an, es handle sich ums Sortiment B . Etwas gelehrter formuliert: Sie lehnt H_0 : “die Schachtel ist Sortiment A ” ab und akzeptiert H_1 : “die Schachtel ist Sortiment B ”. Berechnen Sie die Wahrscheinlichkeit für den Fehler 1. und den Fehler 2. Art.
- 9–8 Willy Würfel besass, wie aus Aufgabe 3–32 bekannt, einen Würfel, bei dem die Wahrscheinlichkeit für eine Sechs 30% betrug. Er hatte aber noch einen andern, bei dem die Sechs eine Wahrscheinlichkeit von nur 10% hatte. Seine Erben konnten die beiden Würfel nicht unterscheiden. Sie beschlossen deshalb, einen der Würfel 15-mal zu werfen. Sollten dabei zwei oder weniger Sechsen fallen, so würden sie davon ausgehen, es handle sich um den 10%-Würfel. Formulieren Sie zu diesem Experiment die Null- und die Alternativhypothese. Berechnen Sie die Wahrscheinlichkeit eines Fehlers 1. bzw. 2. Art.
- 9–9 Der Inhalt von 6 Säcken Puderzucker wurde nachgewogen. Man erhielt folgende Gewichte (in Gramm): 495, 502, 505, 498, 490, 500. Ist die Behauptung, dass die Säcke im Durchschnitt

500 g enthielten, haltbar? Testen Sie mit $\alpha = 5\%$.

- 9–10 Auf einer Geburtstagstorte hat es acht Kerzen. Das (offensichtlich frühreife) Geburtstagskind schreibt sich auf, wie lange jede Kerze gebrannt hat und erhält folgende Zeiten (in Minuten):

13, 16, 11, 15, 13, 11, 10, 15.

- a) Auf der Kerzenpackung stand zu lesen: Die mittlere Brenndauer dieser Kerzen beträgt mindestens 15 Minuten. Versuchen Sie, diese Behauptung mit einem statistischen Test zu widerlegen. Arbeiten Sie mit $\alpha = 0.05$.
- b) Angenommen, Sie hätten aufgrund des Tests die Behauptung aus a) widerlegt. Äußern Sie sich zur Frage, ob Sie dies mit absoluter Sicherheit tun dürfen.
- 9–11 Eine Maschine produziert Schrauben, welche im Mittel eine Länge von 50 mm haben sollten. Es besteht der Verdacht, dass sie nicht mehr korrekt eingestellt ist. Eine Stichprobe von 12 Schrauben lieferte folgende Werte (in mm)

49.5, 51.5, 50.0, 50.1, 51.0, 51.2, 49.8, 49.2, 51.7, 50.1, 50.6, 51.3.

Testen Sie die Hypothese, dass die Maschine richtig eingestellt ist, mit einem Signifikanzniveau von a) 5%, b) 10%. Wie erklären Sie allfällig verschiedene Ergebnisse?

- 9–12 Ein Geschäft verkauft Marzipanrollen mit Gewichtsangabe 80 g. Eine Überprüfung von 25 Packungen ergab ein mittleres Gewicht von 79 g mit einer Standardabweichung von 2.6 g.

- a) Testen Sie die Hypothese, dass das mittlere Gewicht dieser Rollen 80 g betrage.
- b) Testen Sie die Frage, ob allenfalls zuwenig Marzipan abgepackt wurde.
- c) Wie würden die Antworten ausfallen, wenn dieselben Masszahlen mit einer Stichprobe vom Umfang 100 ermittelt worden wären?
- 9–13 Neun Versuchspersonen hatten vormittags und nachmittags je einen Test auszuführen. Es ergaben sich die folgenden Resultate:

Person Nr.	1	2	3	4	5	6	7	8	9
Punktzahl vormittags	100	88	99	95	91	101	91	102	96
Punktzahl nachmittags	104	91	102	95	95	100	93	105	96

Prüfen Sie (unter der Annahme, dass die Differenzen der Punktzahlen stetig und normal verteilt sind) nach, ob die Tageszeit einen Einfluss auf die Testresultate hat. Wählen Sie a) $\alpha = 5\%$, b) $\alpha = 1\%$.

- 9–14 Sieben Versuchspersonen führten mit folgendem Ergebnis eine Diät durch:

Person Nr.	1	2	3	4	5	6	7
Gewicht vorher	70.2	55	90.4	66	81.4	62.3	75
Gewicht nachher	68.2	54	86	66	79.9	63.3	74.5

Man möchte mit einem statistischen Test nachprüfen, ob die Diät tatsächlich einen Gewichtsverlust bewirkt. Wählen Sie $\alpha = 5\%$.

- 9–15 Eine Abfüllmaschine für Mehl war so eingestellt, dass das mittlere Abfüllgewicht pro Packung 1000 Gramm betrug. Nach einer Revision wurde eine Stichprobe von 50 Säcken kontrolliert. Man erhielt einen Durchschnitt von 990 Gramm mit einer Standardabweichung von 30 Gramm. Untersuchen Sie mit einem statistischen Test, ob sich das mittlere Abfüllgewicht verändert hat. Erklären Sie, warum Sie ein- bzw. zweiseitig testen. Arbeiten Sie mit einem Signifikanzniveau von 5%.

- 9–16 Wir gehen von der Annahme aus, das Geburtsgewicht von Neugeborenen sei normal verteilt. Aus einer Stichprobe von 40 Neugeborenen wurde ein Mittelwert von 3300 Gramm mit einer Standardabweichung von 500 Gramm bestimmt.
Kann man aufgrund dieser Daten schliessen, dass das mittlere Geburtsgewicht aller Neugeborenen
- mehr als 3200 g beträgt,
 - mehr als 3150 g beträgt?
- Arbeiten Sie mit $\alpha = 5\%$.
- Wir bleiben bei den oben angegebenen Masszahlen und bei $\alpha = 5\%$. Wie gross müsste der Umfang der Stichprobe mindestens sein, damit die Hypothese “das mittlere Geburtsgewicht aller Neugeborenen ist kleiner als 3200 Gramm” verworfen werden kann?
- 9–17 Aus zwei normal verteilten Grundgesamtheiten mit derselben (wenn auch unbekannt) Varianz wurden die folgenden Stichproben entnommen:
- Grundgesamtheit 1: 25, 27, 28, 28, 30, 30.
Grundgesamtheit 2: 23, 24, 24, 25, 25, 26, 28.
- Prüfen Sie mit statistischen Tests die beiden folgenden Behauptungen nach:
- Die Erwartungswerte der beiden Grundgesamtheiten sind gleich.
 - Der Erwartungswert der 1. Grundgesamtheit ist gleich 27.
- Formulieren Sie jeweils die Nullhypothese, und arbeiten Sie mit einem Signifikanzniveau von 5%.
- 9–18 Zwei Diäten wurden an je 6 Versuchspersonen ausprobiert. Diät A ergab Gewichtsabnahmen von 2.5, 1.8, 3.6, 0.5, 2.2 und 1.4 kg, während Diät B Abnahmen von 2.0, 0.8, 0.0, 2.2, 0.1 und 0.3 kg lieferte.
Sie sympathisieren mit der Diät A und möchten statistisch belegen, dass diese Diät grössere Gewichtsabnahmen bringt.
- Testen Sie hier ein- oder zweiseitig?
 - Wählen Sie ein passendes Testverfahren, und führen Sie den Test mit dem Signifikanzniveau $\alpha = 5\%$ durch. Geben Sie die Null- und die Alternativhypothese klar an.
 - Angenommen, das Testergebnis erlaube die Verwerfung der Nullhypothese. Welches Ereignis hat dann eine Wahrscheinlichkeit $\leq \alpha$?
- 9–19 Aus der Feuerwerksproduktion der Hersteller “Aaah!” bzw. “Oooh!” wurden Vulkane auf ihre Brenndauer geprüft. Eine Stichprobe von 8 Stück der Marke “Aaah!” ergab folgende Zeiten (in Sekunden): 50, 57, 57, 60, 60, 62, 64, 70. Von der Marke “Oooh!” wurden 12 Stück getestet. Der Mittelwert der Stichprobe war um 5 Sekunden grösser als jener der Marke “Aaah!”, die Standardabweichung war bei beiden Stichproben dieselbe. Besteht ein Unterschied zwischen den beiden Produkten in Bezug auf die mittlere Brenndauer der gesamten Produktion? Wir nehmen an, die beiden Grundgesamtheiten seien normal verteilt, mit derselben Varianz.
- Testen Sie hier einseitig oder zweiseitig?
 - Formulieren Sie Ihre Null- und Ihre Alternativhypothese.
 - Führen Sie einen statistischen Test zur Beantwortung der eingangs gestellten Frage durch. Wählen Sie ein Signifikanzniveau von $\alpha = 5\%$.
- 9–20 In einem landwirtschaftlichen Versuchsbetrieb wurde der Einfluss eines neuen Düngemittels auf die Getreideproduktion ermittelt. Dazu wurden 24 gleich grosse Parzellen gebildet; 13 davon wurden mit dem neuen Mittel gedüngt und ergaben eine mittlere Ernte von 540 kg pro Parzelle mit einer Standardabweichung $s = 35$ kg. In den 11 konventionell gedüngten Äckern lag der Durchschnitt bei 505 kg mit einer Standardabweichung von 40 kg. Können wir daraus schliessen, dass die Düngung mit dem neuen Mittel eine signifikante Vermehrung des Ertrags bewirkt? a) $\alpha = 1\%$, b) $\alpha = 5\%$.
- 9–21 In der Gemeinde A wurde bei 18 Milchproben ein mittlerer Fettgehalt (pro Liter) von 36 g mit einer Varianz von 16 g^2 festgestellt. In der Gemeinde B dagegen ergaben 10 Proben einen

mittleren Fettgehalt von 39 g mit einer Varianz von 9 g². Kann man sagen, die Milch aus A habe generell einen geringeren Fettgehalt? Arbeiten Sie mit $\alpha = 5\%$.

- 9–22 Der Zürcher Astronom RUDOLF WOLF (1816-1893) führte über die Dauer von vielen Jahren Experimente mit Würfeln durch. Dabei erhielt er bei 20'000 Würfeln mit einem Würfel die nachstehenden Daten:

Augenzahl	1	2	3	4	5	6
Anzahl Würfe	3246	3449	2897	2841	3635	3932

War dieser Würfel ausgewogen?

- 9–23 Jemand behauptet, die Anzahl der Geburten sei, generell gesehen, gleichmässig auf die vier Quartale des Jahres verteilt. In einem Spital wurde nun die Häufigkeit der Geburten pro Quartal in Prozenten wie folgt registriert:

Quartal	1	2	3	4
Prozentuale Häufigkeit	31%	22%	27%	20%

Prüfen Sie die gegebene Behauptung mit einem statistischen Test nach,

- für den Fall, dass sich die obigen Prozentzahlen auf 200 Geburten beziehen,
- für den Fall, dass sich die obigen Prozentzahlen auf 300 Geburten beziehen.

Schreiben Sie die Nullhypothese auf, und arbeiten Sie mit einem Signifikanzniveau von 5%.

- 9–24 In einem genetischen Experiment erhält man 4 Klassen A, B, C, D. Von der Theorie her erwartet man ein Verhältnis von 1:4:4:16. Bei der Durchführung eines Experiments erhielt man die folgenden absoluten Häufigkeiten:

Klasse	A	B	C	D
Häufigkeit	10	25	35	180

Sind diese Daten mit der Theorie verträglich?

- Bei einem Signifikanzniveau von 5%.
- Bei einem Signifikanzniveau von 1%.

- 9–25 Eine Firma stellt in Papier verpackte Schokoladeneier her. 20% der Produktion sind aus weisser, der Rest aus brauner Schokolade. Diese Eier werden zufällig in Dreierpackungen abgefüllt. Die Zufallsgrösse X bezeichne die Anzahl der weissen Eier in einer Packung.

- Welche Verteilung nehmen Sie für X an?
- Eine Kontrolle von 1000 Packungen ergab Folgendes:

Anzahl weisse Eier	0	1	2	3
Anzahl solcher Packungen	500	400	90	10

Prüfen Sie mit einem Test Ihre in a) getroffene Annahme nach. Wählen Sie $\alpha = 5\%$.

- 9–26 Von 1000 befragten Autofahrern hatten im letzten Jahr 810 keinen Unfall, 170 einen Unfall und 20 zwei oder mehr Unfälle. Man vermutet, dass eine Poisson-Verteilung mit $\mu = 0.2$ vorliegt. Prüfen Sie diese Behauptung mit $\alpha = 0.05$ statistisch nach.

- 9–27 Eine Untersuchung von 320 Familien mit je vier Kindern ergab die unten stehende Aufteilung in Bezug auf das Geschlecht der Kinder:

Anzahl k der Mädchen	0	1	2	3	4
Anzahl Familien mit k Mädchen	42	110	111	48	9

Testen Sie mit einem geeigneten Verfahren die folgenden Hypothesen:

- (A) Die zugehörige Zufallsgrösse $X = \text{Anzahl der Mädchen pro Vierkinderfamilie}$ ist binomial verteilt mit $p = q = 0.5$.
- (B) Die zugehörige Zufallsgrösse $X = \text{Anzahl der Mädchen pro Vierkinderfamilie}$ ist binomial verteilt.

Arbeiten Sie mit einem Signifikanzniveau von 5%.

- 9–28 Im Jahre 1889 (als es noch viele grosse Familien gab!) wurde eine Untersuchung über die Anzahl Knaben in Familien mit 8 Kindern veröffentlicht:

Anzahl Knaben	0	1	2	3	4	5	6	7	8
Häufigkeit	215	1'485	5'331	10'649	14'959	11'929	6'678	2'092	342

Ist die Zufallsgrösse "Anzahl Knaben" binomial verteilt?

- 9–29 In einer Ortschaft ergab eine Zählung der Verkehrsunfälle während 260 Tagen Folgendes: An 89 Tagen ereignete sich kein Unfall, an 97 Tagen je einer, an 43 Tagen je zwei, an 24 Tagen je drei, an 5 je vier Unfälle, und schliesslich gab es je einen Tag mit fünf bzw. sechs Unfällen. Untersuchen Sie mit einem statistischen Test, ob hier eine Poisson-Verteilung angenommen werden darf ($\alpha = 5\%$).

- 9–30 In einer Telefonzentrale wurden während drei Stunden die Anrufe pro Minute gezählt. Man erhielt folgende Daten:

Anzahl k der Anrufe pro Minute	0	1	2	3	4	5	6
Anzahl Intervalle mit k Anrufen	29	42	42	40	22	4	1

Prüfen Sie mit einem statistischen Test nach, ob man aufgrund dieser Stichprobe annehmen darf, die Zufallsgrösse $X = \text{"Anzahl Anrufe pro Minute"}$ folge einer Poisson-Verteilung.

- a) mit Signifikanzniveau $\alpha = 5\%$,
 b) mit Signifikanzniveau $\alpha = 10\%$.

- 9–31 Entstemmen die nachstehenden Daten einer normal verteilten Grundgesamtheit?

Klasse	[25, 27]	(27, 29]	(29, 31]	(31, 33]	(33, 35]	(35, 37]
abs. Häufigkeit	12	12	27	36	39	24

- 9–32 200 Student(inn)en mussten sich sowohl in Mathematik als auch in Physik prüfen lassen. Die Resultate waren:

	Mathematik bestanden	Mathematik nicht bestanden
Physik bestanden	108	24
Physik nicht bestanden	45	23

Besteht zwischen den beiden Prüfungen ein Zusammenhang ($\alpha = 5\%$)?

- 9–33 Eine Umfrage über die bevorzugten Radiostationen ergab folgendes Bild:

- Von den unter 20-jährigen bevorzugten 80 den Sender A, 180 dagegen den Sender B.
 - Von den über 20-jährigen dagegen hörten 67 lieber den Sender A, 93 den Sender B.
- a) Wie würden die Zahlen lauten, wenn die Lieblingssender von der Altersgruppe unabhängig wären?
- b) Prüfen Sie mit einem statistischen Test nach, ob die Bevorzugung des einen oder andern Senders von der Altersgruppe abhängt, und zwar mit $\alpha = 5\%$ und $\alpha = 1\%$.

- 9–34 Bei der Vorbereitung auf eine Prüfung lernten 150 Studierende mit dem Lehrbuch A, 100 mit dem Lehrbuch B. Von den Kandidat(inn)en mit Lehrbuch A bestanden 80%, von jenen mit Lehrbuch B 70%.
- Hängt der Prüfungserfolg vom Lehrbuch ab? Überprüfen Sie diese Frage mit einem statistischen Test (mit Signifikanzniveau $\alpha = 0.05$).
 - Wir ändern die Zahlen etwas ab: Nun haben 225 Studierende das Lehrbuch A, 150 das Lehrbuch B benützt (also je die Hälfte mehr). Die Prozentzahlen sind aber dieselben geblieben (80% bzw. 70%). Wie sieht die Sache jetzt aus?
- 9–35 Eine eher theoretische Aufgabe: In (9.4.2) ist ohne Beweis die Formel für die Dichtefunktion $f(x)$ der χ^2 -Verteilung angegeben worden. Weisen Sie die Gültigkeit dieser Formel für den einfachsten Fall, nämlich für den Freiheitsgrad $\nu = 1$, nach. Tipp: Für die zugehörige Verteilungsfunktion F (und $x \geq 0$) gilt $F(x) = P(X^2 \leq x)$, wobei X der Standard-Normalverteilung folgt. Formen Sie um, und benützen Sie die Beziehung $f = F'$ (Ableitung).