Grundideen der Testtheorie

Die Testtheorie ist ein Spezialgebiet der schliessenden Statistik, bei der es um folgende Grundprobleme geht:

1. Ist eine bestimmte Hypothese mit den Ergebnissen eines Experimentes verträglich? Unterscheiden sich zwei Testergebnisse? (z.B. Sind zwei Rohstoffe bezüglich der Qualität des Endproduktes unterscheidbar? Führt eine bestimmte Investition in den Maschinenpark zu einer besseren Qualität? Hat ein bestimmtes Medikament die gewünschte Wirkung? Führt eine neue Mastmethode zu höheren Erträgen? etc.). In der „Wirklichkeit" sind entsprechende Messwerte nämlich immer von vielen Einflüssen geprägt. Kein Produkt ist genau gleich wie ein anderes. Kein Medikament wirkt auf alle gleich. Keine Mastmethode führt bei Tieren genau zum gleichen Gewicht. Diese Beispiele zeigen: Messwerte streuen und es stellt sich die Frage, wie man angesichts dieser Streuung Qualität oder Wirksamkeit beurteilen kann.

   Wie in der beschreibenden Statistik kann man in der Testtheorie univariate von bivariaten Tests unterscheiden. Bei den univariaten Tests beschäftigen wir uns nur mit einer Variablen (z.B. Körpergewicht). Bei den bivariaten Tests wird der Zusammenhang zweier Variablen untersucht (z.B. Rauchverhalten und Krebsrisiko). Die bivariaten Tests können auf multivariate Tests erweitert werden, mit denen wir den Einfluss mehrere Variablen auf eine Variable untersuchen (z.B. Einfluss von Einkommen, Geschlecht und Ausbildung auf Kaufverhalten bezüglich Schmuck). Wir werden in diesem Kapitel zuerst die univariate Testtheorie ausführen. Später folgen Beispiele für bivariate Tests und am Schluss ein paar kleine Ausblicke in die multivariaten Tests.

2. Schätzen von Verteilungen und von Parametern von Verteilungen (z.B. Schätzen des Erwartungswertes, der Varianz einer Verteilung). In diesem Zusammenhang werden wir ein paar Gütekriterien für Schätzer von Verteilungsparametern erwähnen.

3. Schätzen von Vertrauensintervallen für Teststatistiken: in der Testtheorie treffen wir auf Werte von Zufallsvariablen (Statistiken), welche die „tatsächlichen" Parameter der entsprechenden Verteilungen schätzen sollen. Es ist nützlich zu wissen, in welchem Intervall sich die tatsächlichen Werte mit welcher Wahrscheinlichkeit befinden.

4. Stichprobentheorie: Schätzen von Parametern einer Grundgesamtheit mit Hilfe einer Stichprobe (Anteile, Mittelwerte: Beispiel: Meinungsumfragen, Marktforschung). Auch in diesem Zusammenhang sind wir an Vertrauensintervallen interessiert: z.B. Bei einer Marktforschungsstudie bemängeln 45% der Befragten einen bestimmten Aspekt eines Produktes. Dieser Wert ist als Wert einer Zufallsvariablen zu betrachten. Es wäre nützlich zu wissen, in welchem Intervall sich der tatsächliche Wert mit einer Wahrscheinlichkeit von 95% findet.

Wir werden in der Folge die Grundbegriffe der Testtheorie einführen. Das Problem des Testens von Hypothesen kann man sich an einem einfachen Beispiel klarmachen:

1 Problemlage des Testens

Beispiel 1.1. *Eine Person behauptet, die Fähigkeit zu haben, Walliser Dôle von Waadtländer Gamay in einem Blindtest unterscheiden zu können. Wir überreden die Person, sich einem Test zu unterziehen. Wir nehmen z.B. folgende Anordnung vor. Wir füllen 16 Gläser mit Dôle und mit Gamay von jeweils verschiedenen Produzenten (Die entsprechenden Anteile der Gläser sollten nach Zufallsprinzip festgelegt werden, indem man z.B. mit Excel =runden(zufallszahl()*16;0)*

eingibt). Die Versuchsperson darf die Anteile nicht kennen. Wir setzen voraus, dass die Wahrscheinlichkeit einer zufälligen, richtigen Entscheidung weder vom Vorliegen von Gamay oder Dôle noch von der jeweils vorangehenden Degustation beeinflusst wird). Wir ordnen die Gläser nach dem Zufallsprinzip - indem wir z.B. den Gläsern der Reihe nach Zufallszahlen zuordnen und die Gläser dann der Grösse der Zufallszahlen nach auf einem Tisch anordnen. Wir fordern die Versuchsperson auf, die Gläser der Reihe nach zu degustieren und jeweils festzulegen, ob es sich um Dôle oder Gamay handelt.

Wie gross ist die Wahrscheinlichkeit, zufälligerweise alle Gläser richtig als Dôle oder als Gamay zu klassifizieren? Die Wahrscheinlichkeit, bezüglich eines bestimmten Glases die richtige Entscheidung zu treffen, beträgt bei zufälliger Entscheidung 0.5. *Somit beträgt die Wahrscheinlichkeit, alle Gläser zufälligerweise richtig zu klassifizieren* $0.5^{16} = 0.000015259 = P(X = 16) = \binom{16}{16}0.5^{16}0.5^{0}$. *Gelingt es der Versuchsperson, alle Gläser richtig zu klassieren, ist es zwar immer noch denkmöglich, dass sie in Tat und Wahrheit die beiden Weinsorten blind nicht unterscheiden kann und dass sie zufälligerweise die richtige Wahl getroffen hat. Die Wahrscheinlichkeit dafür ist jedoch sehr klein. Entsprechend ist das Ergebnis schlecht durch Zufall erklärbar und wir sind bereit, die Hypothese zu verwerfen, dass sie die Gläser nicht unterscheiden kann.*

Wir sind aber vermutlich nicht nur bereit, die ausserordentlichen Fähigkeiten der Versuchsperson zu akzeptieren, wenn sie alle Gläser richtig klassiert, sondern auch, wenn sie die meisten Gläser richtig klassiert. So ist es z.B. auch sehr unwahrscheinlich, höchstens ein Glas zufälligerweise falsch zu raten (höchstens ein Glas falsch bedeutet: kein Glas falsch oder nur eines falsch. „Höchstens ein Glas falsch" ist bei 16 Gläsern äquivalent mit „Mindestens 15 Gläser richtig"). Die Wahrscheinlichkeit, zufälligerweise mindestens 15 Gläser richtig zu bestimmen, ist: $P(X \geq 15) = P(X = 15) + P(X = 16) = 1 - P(X \leq 14) = 0.000259399$ *mit* $X \sim B(16, 0.5)$. *Wir würden bei 15 richtig bestimmten Gläsern vermutlich zum Schluss gelangen, dass das Resultat schlecht durch Zufall zu erklären ist. Wir verwerfen die Meinung, die Person habe nur geraten und wir akzeptieren deren ausserordentlichen Fähigkeiten.* ◊

Systematische Analyse des Beispiels

Die Ergebnismenge ist im Beispiel $\Omega = \{$Die Versuchsperson sagt von einem Glas Wein das richtige aus; Die Versuchsperson sagt von einem Glas Wein das falsche aus$\} =: \{$richtig,falsch$\}$. Unter der Hypothese, dass die Versuchsperson die Weine nicht unterscheiden kann, gilt: Graph von $f_\Omega = \{(\text{falsch}, 0.5), (\text{richtig}, 0.5)\}$. Sinnvoll ist die Zufallsvariable (Graph): $X_i = \{(falsch, 0), (richtig, 1)\}$. Wenn wir derart die Summe der Werte der Zufallsvariablen bilden, ergibt sich nämlich die Anzahl der korrekten Tipps. Es gilt: Graph von $f_{X_i} = \{(0, 0.5), (1, 0.5)\}$. Wir haben 16 solcher unabhängiger Zufallsvariablen X_i auf Ω für die gilt: $X_i \sim B(1, 0.5)$. Die Anzahl der richtigen Mutmassungen ist die Summe X dieser 16 identisch verteilten, unabhängigen Bernoulli-Zufallsvariablen. Diese Summe ist eine binomialverteilte Zufallsvariable (= Statistik) $X \sim B(16, 0.5)$. Die folgende Übersicht spiegelt die Situation:

$$\Omega^{16} \xrightarrow{(X_1 \ldots, X_{16})} \bigtimes_{i=1}^{16} \mathfrak{X}_i \xrightarrow{X} \mathbb{N}_{16}$$

($\bigtimes_{i=1}^{16} \mathfrak{X}_i$ = Kartesisches Produkt der Bilder von X_i; $\to$ bezeichnet eine Abbildung; $\mathbb{N}_{16}$ die ersten 16 natürlichen Zahlen inklusive 0).

Wir nennen die Statistik X in diesem Zusammenhang „Teststatistik" und den konkreten Wert, welchen die Zufallsvariable X annimmt „Testwert".

2 Nullhypothese und Alternativhypothese

Wir sind im allgemeinen bereit, eine Hypothese H zu verwerfen, wenn die Wahrscheinlichkeit sehr klein ist, dass *bei der Annahme von* H das spezifische Testergebnis oder ein extremeres Ergebnis

zufällig zustande kommt. Das Ergebnis ist dann ja schlecht durch Zufall zu erklären. Im Beispiel nehmen wir als Hypothese H an, die Versuchsperson könne die Weinsorten nicht unterscheiden. Unter dieser Hypothese ist die Wahrscheinlichkeit sehr klein, dass sie z.B. mindestens 15 richtige Entscheidungen trifft, d.h. 15 oder 16 richtige Entscheidungen trifft. Die Hypothese H, die wir verwerfen, wenn das Testergebnis oder ein extremeres Ergebnis bei der Annahme von H unwahrscheinlich ist, nennen wir künftig Nullhypothese (H_0). Die Negation von H_0 bezeichnen wir mit H_A (= Alternativhypothese). Im Weinbeispiel ist
H_0 : „Die Versuchsperson kann die Weingläser nicht korrekt klassifizieren“
H_A : „Die Versuchsperson kann die Weingläser korrekt klassifizieren“.

Ein Test setzt voraus, dass wir eine Zufallsvariable (= Teststatistik) mit bekannter Wahrscheinlichkeitsverteilung unter der Nullhypothese zur Verfügung haben, mit der wir die Wahrscheinlichkeit des Testergebnisses (oder eines extremeren Ergebnisses) unter der Voraussetzung der Nullhypothese beurteilen können (im Beispiel verwendeten wir die binomialverteilte Zufallsvariable X: $X \sim B(16, 0.5)$).

Bemerkung 2.1. *Die Nullhypothese ist die Hypothese, für die wir eine Teststatistik haben, deren Verteilung wir kennen. Die Nullhypothese besagt, dass der Testwert zur bekannten Verteilung passt oder davon nicht zu stark abweichen, wobei wir die Nullhypothese immer am konkreten Fall inhaltlich ausformulieren. Wenn im obigen Beispiel die Versuchsperson nicht die Fähigkeit hat, die Weingläser korrekt zu klassifizieren, rät sie nur, und die Wahrscheinlichkeit, den Wein im Einzelversuch korrekt zu klassifizieren beträgt 0.5. Damit haben wir eine Verteilung für die Teststatistik, und die Nullhypothese ist, dass die Testperson die Weine nicht richtig klassifizieren kann. Die Alternativhypothese ist die Negation der Nullhypothese.* ◊

Bemerkung 2.2. *Das Wort „Hypothese“ ist aus den griechischen Wörtern „hypó“ (unter, unterhalb) und thésis (Satz, Behauptung) zusammengesetzt. Man übersetzt es in unserem Zusammenhang am besten mit „Vermutung“, wobei solche Vermutungen nicht willkürlich sind, sondern oft wohl begründet. Alle wissenschaftlichen Gesetze sind Hypothesen, manche sind allerdings so gut bestätigt, dass wir an ihnen nicht zweifeln (z.B. Alle Metalle leiten Strom).* ◊

3 Signifikanzniveau

Traditionell legt man ein bestimmtes Niveau für die Verwerfung der Nullhypothese fest. Das Verwerfungsniveau wird *Signifikanzniveau* genannt und gewöhnlich auf $\alpha = 0.05$ oder $\alpha = 0.01$ festgelegt. Fällt der Testwert in einen Bereich, in den bei Annahme der Nullhypothese der Testwert *oder ein extremerer Wert* nur mit einer Wahrscheinlichkeit von weniger als 0.05 oder 0.01 fallen, so verwerfen wir die Nullhypothese. Gewöhnlich spricht man dann von einem *statistisch signifikanten Testwert*. Signifikanzniveaus sind jeweils vor einem Test festzulegen. Oft wird die Signifikanz von Testwerten mit einen Stern für das $0.05-$Niveau, mit zwei Sternen für das $0.01-$Niveau und mit drei Sternen für das $0.001-$Niveau gekennzeichnet.

Legen wir im obigen Beispiel 0.05 als Signifikanzniveau fest, sind wir bereit, die Nullhypothese zu verwerfen, wenn für den Testwert x des Versuchs bei Gültigkeit der Nullhypothese gilt: $P(X \geq x) < 0.05$. Berechnen wir die entsprechende (umgekehrte), kumulative Wahrscheinlichkeitsverteilung, erhalten wir (s. Tabelle 1):

Somit sind wir bereit, die Nullhypothese zu verwerfen, wenn die Versuchsperson mindestens 12 richtige Zuordnungen vornimmt.

Die Menge der Werte, bei deren Annahme durch die Statistik die Nullhypothese verworfen wird, wird auch „Verwerfungsbereich zum Signifikanzniveau α“ genannt (kurz „$\alpha-$Verwerfungsbereich“). Im obigen Beispiel ist der 0.05-Verwerfungsbereich $\{12, 13, 14, 15, 16\}$. Statt zu sagen, dass wir die Nullhypothese verwerfen, wenn $P(X \geq x) < 0.05$, können wir ebensogut sagten, dass wir die Nullhypothese verwerfen, wenn der Testwert x Element des Verwerfungsbereichs ist. Die Grenze

Wert der ZV x_i	$P(X = x_i)$	$P(X \geq x_i)$
16	1.52588E-05	1.52588E-05
15	0.000244141	0.000259399
14	0.001831055	0.002090454
13	0.008544922	0.010635376
12	0.027770996	0.038406372
11	0.066650391	0.105056763
10	0.122192383	0.227249146

Tabelle 1: Von oben kumulierte Werte einer binomialverteilten Zufallsvariable (n = 16; p = 0.5)

des Verwerfungsbereichs nennen wir „kritische Grenze“. Im obigen Beispiel ist die kritische Grenze 12.

„signifikant“ bedeutet soviel wie „bedeutsam“ oder „wichtig“. Bei uns heisst „signifikant“ „bedeutsam, weil schlecht durch Zufall erklärbar“. Bedeutsamkeit ist damit in einem rein statistischen Sinne aufzufassen. Ob Unterschiede inhaltlich (ökonomisch) bedeutsam sind, ist ein anderes Problem. So kann eine neue Mastmethode bei Schweinen zu einem durchschnittlichen Mehrertrag von 200 g führen. Die Abweichung kann bei genügend grosser Stichprobe durchaus statistisch signifikant sein (= schlecht durch Zufall zu erklären). Sie braucht aber ökonomisch nicht bedeutsam zu sein, wenn die Einführung der neuen Methode in Hinblick auf den Ertrag zu viel kostet. Allerdings kann ein Unterschied nicht ökonomisch bedeutsam sein, wenn er statistisch nicht signifikant ist. Denn dann ist der Unterschied recht gut durch Zufall erklärbar und kann entsprechend mit hoher Wahrscheinlichkeit nicht ökonomisch relevant sein.

4 Zweiseitige und einseitige Tests

Im obigen Beispiel 1.1 können wir festlegen, dass wir die Nullhypothese verwerfen, wenn die Versuchsperson mindestens 12 richtige Entscheidungen trifft. Wir sprechen in diesem Falle von einem *rechtseitigen Test* (s. Abbildung 1).

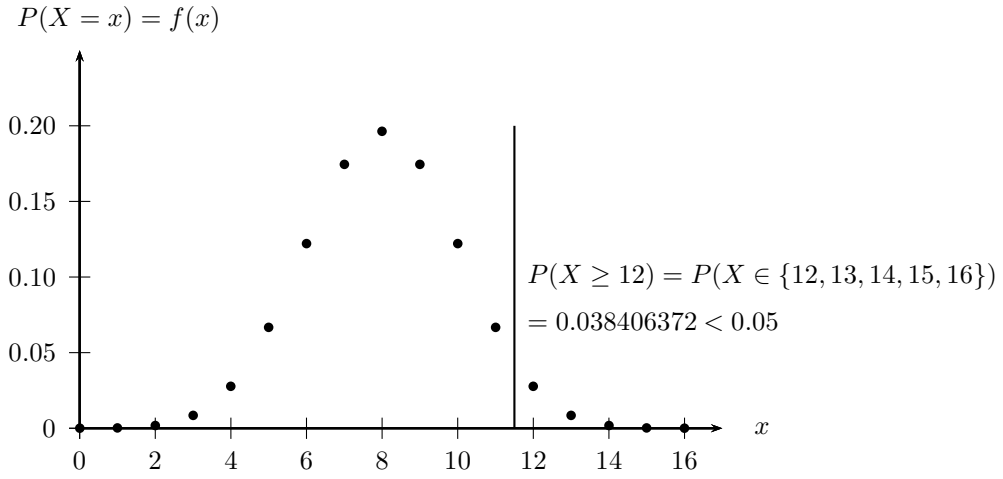

Abbildung 1: Rechtsseitiger Test: $\alpha = 0.05$. Verwerfen der Nullhypothese bei mindestens 12 richtigen Klassierungen; $P(X \geq 12) = 0.038406372 < 0.05$; 0.05-Verwerfungsbereich = $\{12, 13, 14, 15, 16\}$

Es ist auch möglich, auf Grund inhaltlicher Überlegungen den Verwerfungsbereich (der Nullhypothese) auf der linken Seite der Verteilung festzulegen.

Beispiel 4.1. *In einer Hühnerfarm wurden neue Wärmelampen installiert. Mit den alten Wärmelampen gibt es durchschnittlich 20% Ausfälle (80% schlüpfen, 20% nicht). Es wird ein Test mit neuen Wärmelampen gemacht (n = Stichprobengrösse = 50). Dabei gibt es nur 8% Ausfälle. Führen die neuen Wärmelampen zu einer Reduktion der Ausfälle, die schlecht durch Zufall zu erklären sind (d.h. die neuen Wärmelampen wären mit grosser Wahrscheinlichkeit besser als die alten)? (Signifikanzniveau* $\alpha = 0.05$).
H_0 : *mit den neuen Wärmelampen gibt es nicht weniger Ausfälle als mit den alten.*
H_A : *mit den neuen Wärmelampen gibt es weniger Ausfälle als mit den alten (*$\Longrightarrow$ *linksseitiger Test).* $X \sim B(50, 0.2)$. *8% von 50 sind 4. Damit gilt es zu berechnen:* $P(X \leq 4) = \sum_{i=0}^{4} \binom{50}{i} 0.2^i 0.8^{50-i} = 0.018496 < 0.05$. *Wir können die Nullhypothese verwerfen. Die neuen Wärmelampen sind mit hoher Wahrscheinlichkeit besser (s. Abbildung 2).*

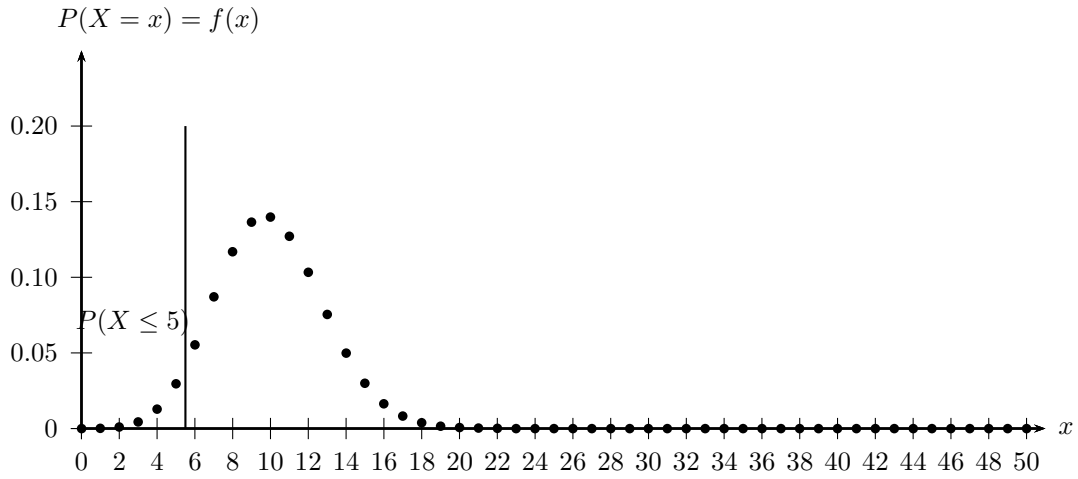

Abbildung 2: linksseitiger Test; $P(X \leq 5) = 0.0480279 < 0.05$ für $X \sim B(50, 0.2)$; 0.05-Verwerfungsbereich $= \{0, 1, 2, 3, 4, 5\}$

Statt zu sagen, dass $P(X \leq 4) = 0.018496 < 0.05$, *und damit die Nullhypothese als genügend unwahrscheinlich zu verwerfen ist, können wir auch hier sagen, dass der Testwert* $4 \in \{0, 1, 2, 3, 4, 5\}$ *und damit Element des Verwerfungsbereichs ist.* $\Diamond$

Rechtsseitige oder linkseitige Tests nennen wir „einseitige Tests“. Neben einseitigen Tests gibt es auch zweiseitige Tests: Nehmen wir an, die Person im obigen Beispiel 1.1 würde 16 falsche Entscheidungen treffen. Es wäre dann wohl sinnvoll anzunehmen, sie könne die Weine unterscheiden, würde diesen aber jeweils falsche Ausdrücke zuordnen. Dies würde uns zu folgenden Hypothesen führen:
H_0 : Die Person kann die Gläser nicht unterscheiden (und klassifiziert sie nicht überwiegend falsch oder nicht überwiegend korrekt).
H_A : Die Person kann die Gläser unterscheiden (und klassifiziert sie überwiegend falsch oder dann überwiegend korrekt)

Entsprechend müssten wir den Verwerfungsbereich auf beiden Seiten der Verteilung festlegen. Bei einem Signifikanzniveau von 0.05 ergibt sich auf beiden Seiten eine Wahrscheinlichkeitsgrenze von 0.025. Betrachten wir wieder unser Weindegustations-Beispiel: Bei beidseitigem Verwerfungsbereich würden wir die Nullhypothese verwerfen, wenn höchstens 3 oder mindestens 13 richtige Entscheidungen getroffen werden (bei $p = 0.5$, ist die Binomialverteilung symmetrisch. Wir können somit durch die Bestimmung einer Grenze mit Hilfe der obigen Tabelle auch die andere Grenze bestimmen). Ob beidseitig oder einseitig getestet werden soll, muss vorgängig zu einem Test auf dem Hintergrund von *inhaltlichen* Überlegungen festgelegt werden (s. Abbildung 3).

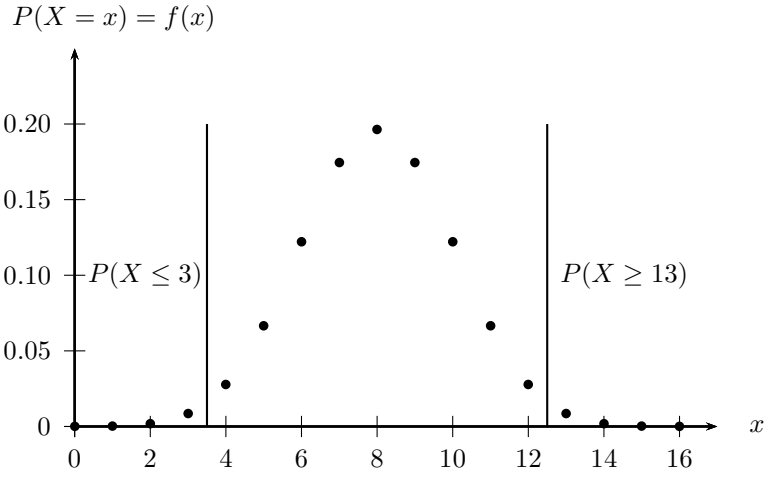

Abbildung 3: Zweiseitiger Test: $\alpha = 0.05$. Verwerfen der Nullhypothese bei höchstens 3 oder mindestens 13 richtigen Klassierungen; $P(X \leq 3) + P(X \geq 13) = 0.010635376 + 0.010635376 = 0.02127 < 0.05$; $0.05-$Verwerfungsbereich $\{0, 1, 2, 3, 13, 14, 15, 16\}$

Bemerkung 4.2. *„Rechts-, links- und zweiseitiger Test" bezieht sich auf die Verwerfungsbereiche und wo diese auf der x-Achse bezüglich der Dichte- oder Wahrscheinlichkeitsfunktion der Verteilung der Teststatistik liegen. Ein Test ist rechtsseitig, wenn der Verwerfungsbereich rechts das Signfikanzniveau α abschneidet. Ein Test ist linksseitig, wenn der Verwerfungsbereich links bei der Dichte- oder der Wahrscheinlichkeitsfunktion α abschneidet. Ein Test ist zweiseitig, wenn der Verwerfungsbereich bei der Dichte- oder Wahrscheinlichkeitsfunktion sowohl rechts als auch links je $\frac{\alpha}{2}$ abschneidet (Der Sprachgebrauch ist allerdings in der Literatur nicht einheitlich. Das Statistikprogramm SPSS braucht z.B. rechts- und linksseitig nicht immer auf diese Art).* ◊

Bemerkung 4.3. *Ob ein Test rechts- oder linksseitig ist, liegt oft an der Formulierung der Hypothesen. So führt die Formulierung „Es hat mit der neuen Brutmaschine mehr ausgeschlüpfte Küken" auf einen rechtsseitigen Test und die Formulierung „Es hat mit der neuen Brutmaschinen weniger nicht ausgebrütete Eier" auf eine linksseitigen Test.* ◊

5 p-Wert

Trifft in einem *rechtsseitigen* Test der Testwert x ein, so wird $P(X \geq x)$ „p-Wert" des Testwertes genannt. Das Testergebnis ist , wie gesehen, signifikant, wenn der p-Wert kleiner als das Signifikanzniveau α ist. Im *linkseitigen Fall* ist der p-Wert des Testwertes x die Wahrscheinlichkeit $P(X \leq x)$. Auch hier ist das Testergebnis signifikant, wenn der p-Wert kleiner als das Signifikanzniveau α ist.

Im *beidseitigen Test* sollte das Testergebnis ebenfalls signifikant sein, wenn die Wahrscheinlichkeit, den Testwert x oder ein extremeres Ergebnis zu haben, kleiner als α ist. Dabei kann der Testwert links oder rechts der Verteilung in einen unter der Nullhypothese unwahrscheinlichen Bereich fallen. Damit verwerfen wir die Nullhypothese genau dann, wenn

$$P(X \leq x) < \frac{\alpha}{2} \text{ oder } P(X \geq x) < \frac{\alpha}{2}$$

- oder anders formuliert - genau dann, wenn

$$\min\{P(X \leq x), P(X \geq x)\} < \frac{\alpha}{2}.$$

Um das folgende Verwerfungsprinzip möglichst einheitlich formulieren zu können, multiplizieren wir beide Seiten mit 2 und erhalten äquivalent

$$2 \cdot \min\{P(X \leq x), P(X \geq x)\} < \alpha.$$

Definieren wir den p-Wert im zweiseitigen Fall durch

$$2 \cdot \min\{P(X \leq x), P(X \geq x)\}$$

können wir für alle drei Fälle festhalten, dass wir die Nullhypothese verwerfen, wenn der p-Wert kleiner als α ist. Zusammenfassend halten wir fest:

Definition 5.1. *Sei x der Testwert.*
Bei einem rechtseitigen Test ist der p-Wert: $P(X \geq x)$
Bei einem linksseitigen Test ist der p-Wert: $P(X \leq x)$
Bei einem zweiseitigen Test ist der p-Wert: $2 \cdot \min\{P(X \leq x), P(X \geq x)\}$ $\Diamond$

Wir betrachten ein Beispiel zum zweiseitigen Fall: wir legen vorgängig zum Weinbeispiel fest, dass wir zweiseitig testen wollen. Die Versuchsperson entscheidet 10 mal richtig.

$$\begin{aligned} P(X &\leq 10) = 0.894943237 \\ P(X &\geq 10) = 1 - F(X \leq 9) = 1 - 0.772\,75 = 0.227\,25 \end{aligned}$$

Damit ist der p-Wert

$$2 \cdot \min\{0.894943237, 0.227\,25\} = 2 \cdot 0.227\,25 = 0.454\,5.$$

Man könnte im zweiseitigen Fall statt $2 \cdot \min\{P(X \leq x), P(X \geq x)\}$ mit α zu vergleichen, wie die obige Argumentation zeigt, auch das Signifikanzniveau halbieren und dieses direkt mit $P(X \leq x)$ und mit $P(X \geq x)$ vergleichen - in der Praxis werden wir das oft auch tun. Wir könnten dann allerdings das folgende Verwerfungsprinzip nicht so einfach formulieren.

Definition 5.2. ***Verwerfungsprinzip**: Wenn*
(1) x eintrifft und
(2) $p-$Wert (von x) $< \alpha$ unter der Voraussetzung der Geltung von H_0
dann verwerfen wir H_0 (α ist das Signifikanzniveau; im Allgemeinen $\alpha = 0.05$ oder $\alpha = 0.01$) $\Diamond$

Alternativ könnte man das Verwerfungsprinzip auch mit Hilfe des Verwerfungsbereichs definieren: Wenn (1) x eintrifft und
(2) $x \in \alpha-$Verwerfungsbereich V_α, dann verwerfen wir H_0 (Der $\alpha-$Verwerfungsbereich wird im rechtsseitigen Fall bestimmt als $V_\alpha = \{y \mid P(X \geq y) < \alpha\}$; im linksseitigen Fall als $V_\alpha = \{y \mid P(X \leq y) < \alpha\}$; im zweiseitigen Fall als $V_\alpha = \{y \mid P(X \leq y) < \frac{\alpha}{2}\} \cup \{y \mid P(X \geq y) < \frac{\alpha}{2}\}$
Wir werden künftig im Allgemeinen die erste Variante verwenden. Diese hat den Vorteil, dass durch die Angabe des p-Wertes klar wird, ob ein Testergebnis knapp oder deutlich signifikant wird.

6 Fehler erster und zweiter Art

Wir nehmen an, dass wir einen Testwert erhalten, für dessen $p-$Wert gilt: $p-\text{Wert} < \alpha$. Es ist zwar recht unwahrscheinlich, aber trotzdem möglich, dass die Nullhypothese wahr ist. In diesem Falle verwerfen wir die Nullhypothese fälschlicherweise. Ein Verwerfen der richtigen Nullhypothese wird „*Fehler erster Art*“ genannt (Wenn man einen Test mit einem Signifikanzniveau von 0.05 und einer richtigen Nullhypothese 100 mal wiederholt, so werden wir mit hoher Wahrscheinlichkeit einige signifikante Resultate erhalten, nämlich ca. 5 auf 100. Die Anzahl der signifikanten Resultate ist dabei als Wert einer Zufallsvariable zu betrachten). Handkehrum ist es möglich, dass wir einen

nicht-signifikanten $p-$Wert erhalten, obwohl die Nullhypothese nicht zutrifft. Diesen Fehler nennt man „*Fehler zweiter Art*“. Die Fehlerwahrscheinlichkeiten für die beiden Fehlertypen hängen vom Signifikanzniveau α ab. Wird das Signifikanzniveau bei diskret verteilten Teststatistiken genügend stark gesenkt, so sinkt die Fehlerwahrscheinlichkeit erster Art und die Fehlerwahrscheinlichkeit zweiter Art wird erhöht.

Wir betrachten dazu zuerst das Beispiel einer binomialverteilten Statistik (Weinbeispiel). Nehmen wir an, die Person könne die Weine korrekt klassifizieren, wobei die Wahrscheinlichkeit einer korrekten Klassifikation 0.9 betrage. Die tatsächliche Verteilung kennen wir i.A. nicht. Wir nehmen hier einfach eine an, um Überlegungen zum Fehler 1. und 2. Art anzustellen. Tragen wir die Wahrscheinlichkeitsfunktionen der Teststatistik unter der Nullhypothese und der Hypothese ab, dass die Person mit $p = 0.9$ korrekt klassifiziert, in einer Graphik ab, so erhalten wir (s. Abbildung 4)

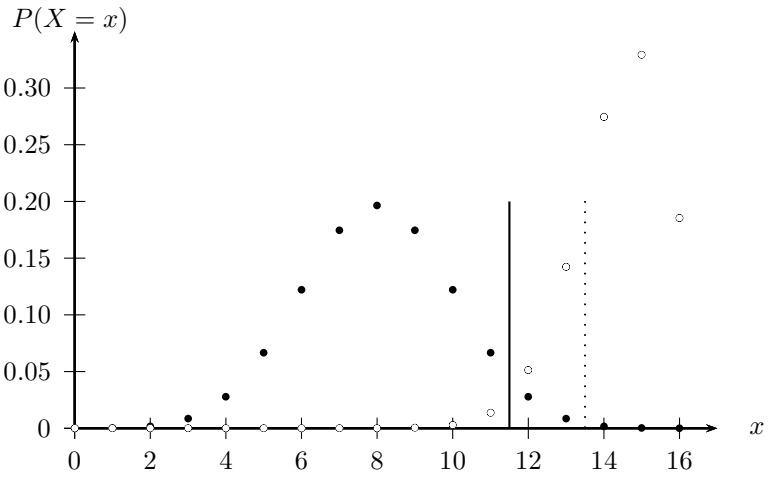

Abbildung 4: Fehler 1. und 2. Art für einem rechtsseitigen Test bei $\alpha = 0.05$ und $\alpha = 0.01$. Unter der Nullhypothese ist die Statistik $B(16, 0.5)$ verteilt (fette Punkte). Die tatsächliche Wahrscheinlichkeitsfunktion sei durch $B(16, 0.9)$ gegeben (nicht ausgefüllte Punkte).

Die Verkleinerung des Signifikanzniveaus (von 0.05 auf 0.01) bewirkt eine Verminderung der Fehlerwahrscheinlichkeit erster Art und eine Vergrösserung der Fehlerwahrscheinlichkeit zweiter Art. Bei $\alpha = 0.05$ (durchgezogene Linie) beträgt die Fehlerwahrscheinlichkeit erster Art $P(X \geq 12) = 0.038406372$ für $X \sim B(16; 0.5)$ und die Fehlerwahrscheinlichkeit zweiter Art $P(X \leq 11) = 0.017003998$ für $X \sim B(16; 0.9)$. Bei $\alpha = 0.01$ (punktierte Linie) beträgt die Fehlerwahrscheinlichkeit erster Art $P(X \geq 14) = 0.002090454$ für $X \sim B(16; 0.5)$ und die Fehlerwahrscheinlichkeit zweiter Art $P(X \leq 13) = 0.21075066$ für $X \sim B(16; 0.9)$.

Bei stetig verteilten Teststatistiken beträgt die Fehlerwahrscheinlichkeit erster Art genau α. In der folgenden Graphik betrachten wir eine standardnormalverteilte Teststatistik (dargestellt durch die Dichtefunktion der Zufallsvariable unter der Nullhypothese). Daneben betrachten wir eine $X \sim N(3, 1)$ verteilte Zufallsvariable (dargestellt durch die Dichtefunktion der tatsächlichen Verteilung der untersuchten Grösse; die tatsächliche Verteilung kennen wir auch hier gewöhnlich nicht. Wir nehmen sie einfach an; s. Abbildungen 5 und 6). Man sieht, dass die Reduktion der Fehlerwahrscheinlichkeit 1. Art nur durch eine Erhöhung der Fehlerwahrscheinlichkeit 2. Art erkauft werden kann - und umgekehrt.

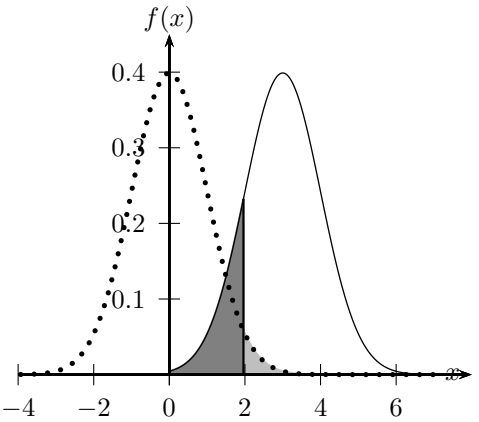

Abbildung 5: gepunktete Dichtefunktion = Dichtefunktion unter der Verteilung der Nullhypothese; durchgezogene Dichtefunktion = Dichtefunktion der „wirklichen“ Verteilung; rechtsseitiger Test mit Verwerfungsbereich = $]1.96, \infty[$; hellere graue Fläche = Fehlerwahrscheinlichkeit 1. Art (= 0.025);dunklere graue Fläche = Fehlerwahrscheinlichkeit 2. Art (= 0.14916995)

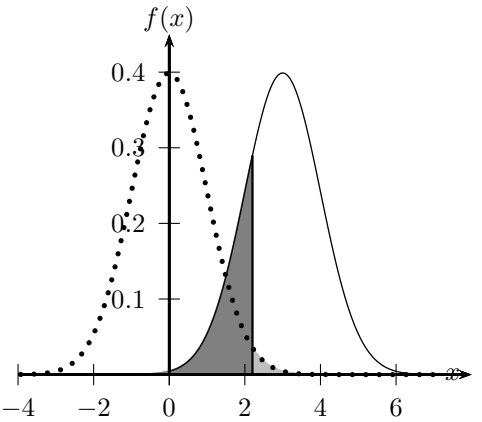

Abbildung 6: gepunktete Dichtefunktion = Dichtefunktion unter der Verteilung der Nullhypothese; durchgezogene Dichtefunktion = Dichtefunktion der „wirklichen“ Verteilung; rechtsseitiger Test mit Verwerfungsbereich = $]2.2, \infty[$; hellere graue Fläche = Fehlerwahrscheinlichkeit 1. Art (= 0.014);dunklere graue Fläche = Fehlerwahrscheinlichkeit 2. Art (= 0.211855399)

Auf Grund von aussermathematischen Überlegungen zu den beiden Fehlertypen kann man die Festlegung des Signifikanzniveaus unter Umständen begründen, indem man die Kosten beim Auftreten der beiden Fehler analysiert. Wir betrachten als Beispiel die Verwendung eines Medikamentes mit starken, negativen Nebenwirkungen. Führt uns ein Test fälschlicherweise dazu, die Wirksamkeit des Medikamentes zu postulieren, so werden Patienten den Nebenwirkungen ausgesetzt, ohne dass das Medikament wirkt. Dies sollte man vermeiden. Entsprechend muss der Fehlerwahrscheinlichkeit erster Art reduziert werden und das Signifikanzniveau ist klein anzusetzen (kleines Alpha). Hat ein Medikament demgegenüber keine bekannten Nebenwirkungen, ist es weniger wichtig, die Verabreichung des Medikamentes an eine möglichst klar bestätigte Wirksamkeit zu binden. Das Signifikanzniveau kann höher liegen (Alpha kann grösser sein).

Beispiel 6.1. *Weiteres Beispiel für Fehler erster und zweiter Art: In einer Population sei der Anteil von Personen, die eine Schokolade gut finden* 0.4*. Die Firma verändert die Rezeptur, um zu sehen, ob der Anteil an Personen, welche die Schokolade gut finden, gesteigert werden kann. Es wird eine Zufallsstichprobe (Stichprobengrösse* $n = 100$*) gezogen, um die Anteilsentwicklung zu überprüfen. Ein Fehler erster Art liegt vor, wenn sich der Anteil an Personen, welche die Schokolade mit neuer Rezeptur gut finden, in der Population nicht ändert, die Firma aber zufälliger Weise einen Stichprobe gezogen hat, welche z.B. einen Anteil von* 0.5 *solcher Personen aufweist. Bei einem Signifikanzniveau von* 0.05 *wird die Nullhypothese, dass sich der Anteil nicht erhöht hat, nämlich in diesem Fall verworfen (es gilt* $P(X \geq 50) = 0.027099198 < 0.05$ *mit* $X \sim B(100, 0.4)$*). Ein Fehler zweiter Art liegt vor, wenn sich der Anteil in der Population z.B. auf* 0.5 *erhöht hat, die Firma aber zufälliger Weise eine Stichprobe zieht, in der der Anteil* 0.4 *ist. Die Nullhypothese wird nicht verworfen, obwohl man sie hätte verwerfen sollen.* ◊

Beispiel 6.2. *Noch ein ökonomisches Beispiel für die Problematik der Festlegung des Signifikanzniveaus: Wir nehmen an, eine Warenlieferung werde mit Hilfe eines Testes auf die Erfüllung der minimalen, vertraglich abgemachten Qualitätsstandards untersucht. Laut diesen dürfen höchstens 5% der Produkte einer Warenlieferung defekt sein. Bei einer Warenlieferung von 200 Produkten dürfen entsprechend nur 10 Produkte defekt sein. Da eine Untersuchung aller Produkte zu kostspielig ist, ziehen wir eine Zufallsstichprobe von 30 Produkten, um diese auf ihren Zustand*

zu untersuchen. Da die untersuchte Gesamtheit relativ klein ist und wir defekte Objekte nicht zurücklegen möchten, verwenden wir als Teststatistik eine hypergeometrisch verteilte Zufallsvariable. Wir nehmen an, in der Stichprobe befänden sich 5 defekte Produkte. Erfüllt die Lieferung die Qualitätsanforderungen?
H_0 : Die Lieferung weist nicht mehr defekte Objekte auf als erlaubt.
H_A : Die Lieferung weist mehr defekte Objekte aus als erlaubt (rechtsseitiger Test).
Wie ist das Signifikanzniveau festzulegen? Bei einem kleinen α wird die Wahrscheinlichkeit, eine Warenlieferung zu akzeptieren, die den Qualitätsstandards nicht genügt, vergrössert (= Wahrscheinlichkeit, Fehler zweiter Art zu begehen, wird vergrössert). Dies ist nicht im Interesse des Belieferten. Andererseits wird dadurch die Wahrscheinlichkeit, eine Warenlieferung zurückzuweisen, die den Ansprüchen genügt, verkleinert (= Wahrscheinlichkeit, Fehler erster Art zu begehen, wird verkleinert). Dies liegt im Interesse des Lieferanten. Die Festlegung des Signifikanzniveaus liegt damit im Schnittpunkt widerstrebender Interessen. Es liegt an den Vertragspartnern, den möglichen Schaden von Fehleinschätzungen überzeugend ins Feld zu führen, um ein für sie möglichst günstiges Signifikanzniveau auszuhandeln.

Um ein weiteres Beispiel für die Durchführung eines Tests zu haben, führen wir das eben eingeführte Beispiel weiter. Wir nehmen an, die Vertragspartner hätten sich auf ein Signifikanzniveau von 0.01 *geeinigt. Wir berechnen $P(X \geq 5)$ mit $X \sim H(200, 10, 30)$.* $P(X \geq 5) = \sum_{i=5}^{10} \frac{\binom{190}{30-i}\binom{10}{i}}{\binom{200}{30}} = 0.008042\,6 < 0.01$. *Die Nullhypothese kann verworfen werden. Es ist unwahrscheinlich, dass unter der Nullhypothese mindestens 5 defekte Objekte in der Stichprobe auftauchen. Die Lieferung entspricht den Qualitätsansprüchen nicht.* ◊

Bemerkung 6.3. *Auch im Beispiel 6.2 kann man die Situation mit einer Übersicht der folgenden Art darstellen, da man eine hypergeometrisch verteilte Zufallsvariable als Summe von bernoulliverteilten Zufallsvariablen betrachten kann (s. Seite **??**).*

$$\Omega^{30} \xrightarrow{(X_1,\ldots,X_{30})} \bigtimes_{i=1}^{30} \mathfrak{X}_i \xrightarrow{X} \mathbb{N}_{10}$$

◊

7 Macht eines Testes

Als Macht, Güte oder Trennschärfe eines Testes wird die Wahrscheinlichkeit bezeichnet, die falsche Nullhypothese zu verwerfen. Die Macht des Testes ist also die Wahrscheinlichkeit, eine korrekte Alternativhypothese zu akzeptieren. Es handelt sich um die Gegenwahrscheinlichkeit zur Fehlerwahrscheinlichkeit zweiter Art. Beträgt die Fehlerwahrscheinlichkeit zweiter Art z.B. 0.2, beträgt die Macht des Testes 0.8.

Es folgt ein Beispiel für einen Test, der auf einer poissonverteilten Teststatistik beruht:

Beispiel 7.1. *Einer Firma wird vorgeworfen, die Asbest-Grenzwerte in ihren Hallen nicht einzuhalten. Man einigt sich auf folgenden Test: Es werden* 10 *mal jeweils ein Liter Luft untersucht und die Anzahl Asbestfasern gezählt. Der offizielle Grenzwert liegt bei höchstens 3 Fasern pro Liter. Die Nullhypothese ist: „Die tatsächliche Anzahl liegt nicht über dem Grenzwert". Die Alternativhypothese ist: „Die Anzahl liegt über dem Grenzwert" ($\Longrightarrow$ rechtsseitiger Test). Wir erhalten für die Proben folgende Stichprobenresultate:* $1; 3; 5; 7; 2; 6; 4; 3; 2; 1$ *(Anzahl Fasern pro Liter). Unter der Nullhypothese können wir von 10 poisson-verteilten Zufallsvariablen X_i ausgehen, mit $E(X_i) = 3$. Die Summe dieser 10 unabhängigen und identisch verteilten Zufallsvariablen ist dann $X \sim P(30)$. (siehe Sätze zur Summe von Poissonverteilungen im Kapitel „Poissonverteilung"; Ergebnismenge Ω = mögliche Anzahl von Fasern, abzählbar unendlich viele Elemente!; $\mathfrak{X}_i = \mathbb{N}$). Wir gelangen also wieder zur Übersicht:*

$$\Omega^{10} \xrightarrow{(X_1,\ldots,X_{10})} \bigtimes_{i=1}^{10} \mathfrak{X}_i \xrightarrow{X} \mathbb{N}$$

Die Teststatistik X nimmt im vorliegenden Fall den Testwert $1+3+5+7+2+6+4+3+2+1 = 34$ *an. Durchschnittlich liegt das Resultat somit über dem Grenzwert. Es stellt sich die Frage, ob der Testwert die Nullhypothese wiederlegt. Der* $p-$*Wert ist* $P(X \geq 34)$*. Wir berechnen* $P(X \leq 33) = 0.744\,45$ *für* $X \sim P(30)$*. Der p-Wert ist damit* $1 - 0.744\,45 = 0.255\,55 = P(X > 33) = P(X \geq 34) \geq 0.05$ *und er liegt über dem Signifikanzniveau. Die Nullhypothese kann nicht verworfen werden (Da* $P(X \leq 38) = 0.93516$ *und* $P(X \leq 39) = 0.953746962$ *für* $X \sim P(30)$*, müsste somit die Nullhypothese erst bei* 40 *und mehr Fasern verworfen werden (*$\alpha = 0.05$*))* $\Diamond$

Die eingeführten Grundbegriffe (einseitiger und zweiseitiger Test, Verwerfungsniveau = Signifikanzniveau, Verwerfungsbereich, Nullhypothese, Alternativhypothese, p-Wert, Teststatistik, Testwert, Stichprobe, Fehler erster und zweiter Art) werden auch für die später behandelten Tests, die eine stetige Verteilung voraussetzen, verwendet. Es ist wichtig, diese Begriffe verstanden zu haben, da sie in der Folge immer wieder auftauchen. Entsprechend sollten sie wiederholt werden, sobald an weiteren Beispielen die Vorgehensweise des Testens deutlicher geworden ist.

7.1 Übungen

1. Eine Unternehmung fertigt mit einer Maschine täglich elektronische Bauteile. Darunter fallen durchschnittlich 3% defekte an. Es wird eine neue Maschine getestet. Aus der Produktion von 500 Objekten der neuen Maschine wird eine Zufallsstichprobe von 50 Produkten gezogen. Es befinden sich 2 defekte darunter. Ist die neue Maschine besser? ($\alpha = 0.05$)

2. Eine Unternehmung fertigt mit einer Maschine täglich elektronische Bauteile. Darunter fallen durchschnittlich 4% defekte an. Es wird eine neue Maschine getestet. Aus der Produktion der neuen Maschine wird eine Zufallsstichprobe von 75 Produkten gezogen. Es befindet sich 1 defektes darunter. Ist die neue Maschine besser? ($\alpha = 0.05$)

3. Eine Firma wird von der Gewerkschaft beschuldigt, weniger als durchschnittlich für die Sicherheit ihrer Mitarbeiter vorzukehren, da die Unfallquote im Jahre 2000 10 auf 1000 betrug, während sie in vergleichbaren Firmen nur 5 auf 1000 betrug. Kann die vorliegende Unfallhäufigkeit auch durch Zufall erklärt werden? (Signifikanzniveau 0.05)

4. Bei einer Lieferung von Rohstoffteilen dürfen laut Liefervertrag durchschnittlich nur 0.5% mangelhaft sein. Dabei wird der folgende Test verwendet, um die Einhaltung der Lieferbedingungen zu überprüfen: Es werden 20 Teile zufällig gezogen (ohne Zurücklegen). Wir nehmen an, bei einer solchen Stichprobe und einer Lieferung von 200 Teilen seien 2 Bauteile fehlerhaft gewesen. Entspricht die Lieferung den Lieferbedingungen? ($\alpha = 0.01$)

5. Bei einer Lieferung von Zwischenprodukten dürften laut Liefervertrag durchschnittlich nur 1% mangelhaft sein. Dabei wird der folgende Test verwendet, um die Einhaltung der Lieferbedingungen zu überprüfen: Es werden 20 Teile zufällig gezogen (ohne Zurücklegen). Wir nehmen an, bei einer solchen Stichprobe und einer Lieferung von 500 Teilen seien 2 Bauteile fehlerhaft gewesen.
   a) Entspricht die Lieferung den Lieferbedingungen? ($\alpha = 0.05$)
   b) Welche Interessen kommen bei der Festlegung des Signifikanzniveaus ins Spiel? (Ist der Lieferant oder der Belieferte an einem möglichst hohen Signifikanzniveau interessiert? Welche Gewinne und Verluste fahren die beiden Beteiligten beim Vorliegen eines Fehlers erster oder zweiter Ordnung ein?).

6. In einer Zeitung steht, die Unfälle mit Todesfolge hätten in der Schweiz massiv zugenommen, da sie von 1000 auf 1100 gestiegen sind. Testen Sie!

7. Eine Person gibt vor, hellseherische Kräfte zu haben. Insbesondere will sie voraussagen können, ob sich bei einem Münzwurf Kopf oder Zahl ergibt. Denken Sie sich einen statistischen Test aus, um diese Behauptung zu überprüfen (Legen sie dabei H_0 und α fest, sowie ob einseitig oder zweiseitig getestet werden soll).

8. Eine Firma beschwert sich bei einem Lieferanten über unterschiedliche Qualitäten von Lieferungen. Dabei könnten von Auge deutlich zwei Qualitäten unterschieden werden. Der Lieferant bestreitet die Behauptung. Man einigt sich auf folgenden Test: 10 Personen untersuchen je 10 Stichproben, die aus den (angeblich oder wirklich) unterschiedlichen Qualitäten stammen. Als Signifikanzniveau wird 0.01 festgelegt. Man testet beidseitig. Legen Sie die Nullhypothese und die Alternativhypothese fest und berechnen x_i, so dass der p-Wert von $x_i \leq 0.01$.

9. Im Boulvard-Gratis-Blatt „20 minuten“ steht am 11. November 2003 auf der ersten Seite der Titel „Jugendkriminalität stieg um 4 Prozent“. Unter diesem Titel werden folgende Zahlen geliefert: Im Jahr 2002 sind 13'000 Minderjährige wegen Straftaten verurteilt worden. Das zeigt eine vom Bundesamt für Statistik am 10. November 2003 veröffentlichte Erhebung. Seit der Statistik 1999 haben die Verurteilungen damit um rund 500 Fälle oder vier Prozent zugenommen. Ist der Unterschied der beiden Jahre signifikant? ($\alpha = 0.05$).

10. (Beispiel aus dem Tages-Anzeiger, 29. Juli 04, S. 26). Bei der Untersuchung von Invaliden-Renten-Aspiranten wird unter anderem auch Statistik eingesetzt. Der Neuropsychologe Thomas Merten untersuchte einen 22-Jährigen, der vorgab, nach einem Schädel-Hirn-Trauma nicht mehr rechnen zu können. Merten liess den Patienten 40 einfache Rechenaufgaben machen, z.B. 3 + 3 und bot ihm jeweils zwei Lösungen an: die richtige (im Beispiel 6) und eine falsche, die nahe bei der richtigen lag (im Beispiel etwa 7). Der Patient entschied sich 11 mal für die richtige Lösung, 29 mal für die falsche. Merten schliesst aus dem Ergebnis, dass der Patient simuliert, denn der Begutachtete wählte mehr Falsche als eine Person, die wirklich nicht rechnen kann und damit zufällig zwischen wahr und falsch unterscheidet. Führen Sie den Test durch. ($\alpha = 0.05$)

11. Kassasturz SF, 6. November 2012. Ein Pendler, der behauptet, ein mit seinen geheimen Methoden von magnetischen Wellen befreites Handy unter anderen Handis mit Hilfe von Pendeln identifizieren zu können, macht in der Sendung folgenden Test: 10 mal hat er die Möglichkeit, Handis, die unter 10 weissen Plastikschüsseln liegen, zu bependeln. Er weiss, dass es in jeder Testserie genau ein gemäss seiner Methode von magnetischen Wellen befreites Handy hat. Er hat 3 Treffer. Wie hoch ist die Wahrscheinlichkeit, mindestens 3 Treffer zu haben, unter der Nullhypothese, dass er das strahlenfreie Handy nicht erpendeln kann? Das Schweizer Fernsehen gibt an, dem Pendler eine faire Chance zu geben, seine Fähigkeit unter Beweis zu stellen: wenn er 8 richtige Treffer habe, dann würde ihm die Preissumme von 10 000 Franken zugestanden. Wie hoch ist die Wahrscheinlichkeit, unter der Nullhypothese mindestens 8 richtige Treffer zu haben ? Wie beurteilen Sie die Testanlage des SF?

12. Denken Sie sich je eine betriebswirtschaftlich relevante Problemlage aus, die Sie mit Hilfe eines Binomial- und eines Poisson-Testes lösen können. Erfinden Sie eine solche Problemlage auch für einen hypergeometrischen Test.

7.2 Lösungen

1. H_0 : Die neue Maschine produziert nicht weniger Ausschuss.
H_A : Die neue Maschine produziert weniger Ausschuss ($\Longrightarrow$ linksseitiger Test). Als Teststatistik verwenden wir eine hypergeometrisch verteilte Zufallsvariable X mit $X \sim H(500, 50, 15)$ (denn die untersuchte Gesamtheit ist klein und wir möchten defekte Objekte nicht zurücklegen. 3% von 500 sind 15); $P(X \leq 2) = \sum_{i=0}^{2} \frac{\binom{485}{50-i}\binom{15}{i}}{\binom{500}{50}} = 0.818\,32 > 0.05$. Die neue Maschine ist

nicht besser (Dies ist auch nicht verwunderlich: 3% von 50 sind $0.03 \cdot 50 = 1.5$. Es hat somit sogar mehr defekte in der neuen Produktion als im Durchschnitt bei der alten Maschine).

2. H_0 : Die neue Maschine produziert nicht weniger Ausschuss.
   H_A : Die neue Maschine produziert weniger Ausschuss ($\Longrightarrow$ linksseitiger Test). Als Teststatistik verwenden wir eine binomialverteilte Zufallsvariable, da die Grösse der untersuchten Gesamtheit (Neuproduktion) nicht bekannt ist. Ist diese klein, können wir gezogene Stücke zurücklegen, wobei wir defekte jeweils kennzeichnen können, damit wir sie nachher entfernen können. $X \sim B(75, 0.04)$. Wir berechnen $P(X \leq 1) = \sum_{i=0}^{1} \binom{75}{i} 0.04^i (1 - 0.04)^{75-i} = 0.193\,09 > 0.05$. Die neue Maschine ist nicht besser.

3. H_0 : Die Unfallquote der Firma ist nicht grösser als die durchschnittliche Unfallquote. H_A : Die Unfallquote der Firma ist grösser als die durchschnittliche Unfallquote ($\Longrightarrow$ rechtsseitiger Test).
   Wir verwenden als Teststatistik eine poisson-verteilte Zufallsvariable. Wir berechnen somit $P(X \geq 10)$ für $X \sim P(5)$. $P(X \geq 10) = 1 - P(X < 10) = 1 - 0.968171943 = 0.031828 < 0.05$. Somit ist die Abweichung von der durchschnittlichen Unfallquote als signifikant zu bezeichnen. Die Firma sollte etwas für den Unfallschutz unternehmen.

4. H_0 : In der Lieferung hat es nicht mehr defekte Objekte als erlaubt.
   H_A : In der Lieferung hat es mehr defekte Objekte als erlaubt ($\Longrightarrow$ rechtsseitiger Test)
   $\alpha = 0.01$
   Verzichten wir auf Zurücklegen, wird die Hypergeometrische Verteilung angewendet. Da 0.5% von 200 1 ist, kann es in keiner Stichprobe 2 haben. Somit ist $P(X \geq 2) = 0$. Die Nullhypothese muss verworfen werden. Verwenden wir mit Zurücklegen einen Binomialtest erhalten wir folgendes Resultat:
   Wir berechnen: $P(X \geq 2) = 1 - P(X < 2)$ für $X \sim B(20, 0.005)$ und erhalten $1 - 0.995526106 = 0.004473894 < 0.01$.
   Die Nullhypothese ist widerlegt.

5. H_0 : Die Lieferung enthält nicht mehr defekte Zwischenprodukte als erlaubt.
   H_A : Die Lieferung enthält mehr defekte Zwischenprodukte als erlaubt (rechtsseitiger Test)
   1% auf 500 macht 5. Wir verwenden einen Hypergeometrischen Test (kleine untersuchte Gesamtheit!!!). Wir berechnen $P(X \geq 2)$ für $X \sim H(N, n, M) = H(500, 20, 5)$ (N für Anzahl untersuchte Gesamtheit; M für Anzahl Defekte in der Gesamtheit; n = Stichprobengrösse; die Reihenfolge der Angabe der Parameter kann von Buch zu Buch verschieden sein. Bei Excel wird z.B. eine andere Reihenfolge gewählt).

Werte x_i der ZV X	$P(X = x_i)$
$X \sim H(500, 20, 5)$	
2	0.013634848
3	0.000513446
4	9.11126E-06
5	6.07417E-08
$P(X \geq 2) = \sum_{i=2}^{5} P(X_i = x_i) =$	0.014157466

Tabelle 2: Berechnung des p-Wertes bei der Hypergeometrischen Verteilung; im Beispiel $X \sim H(500, 20, 5)$

Da $0.014157466 < 0.05$ ist die Abweichung signifikant. Die Lieferung entspricht den Lieferbedingungen nicht.

6. H_o : Die Unfallzahlen sind im zweiten Jahr nicht höher. H_A : Die Unfallzahlen im zweiten Jahr sind höher ($\Longrightarrow$ rechtsseitiger Test). Da sowohl 1000 als auch 1100 als Werte einer Zufallsvariablen zu betrachten sind, wäre es hier ungünstig, so wie im Beispiel 3 vorzugehen. Wir können jedoch die Situation so deuten, dass unter der Nullhypothese ein Unfall mit gleicher Wahrscheinlichkeit im ersten Jahr erfolgt oder im zweiten Jahr. Laut Nullhypothese ist dann die Wahrscheinlichkeit, dass ein Unfall im ersten Jahr erfolgt $= 0.5$. Wir verwenden entsprechend eine Binomialverteilung: Wir berechnen $P(X \geq 1100)$ für $X \sim B(2100, 0.5)$. Mit der Näherungsformel für die Normalverteilung erhalten wir: $1 - \Phi\left(\frac{1100-(2100\cdot 0.5)}{\sqrt{2100\cdot 0.5\cdot 0.5}}\right) = 1 - \Phi(2.182\,2) = 1 - 0.985452664 = 0.014547 < 0.05$.
Somit ist das Resultat auf einem Signifikanzniveau von 0.05 signifikant, auf einem Niveau von 0.01 nicht signifikant. Mit der Binomialverteilung erhält man: 0.015360354 (=1-BINOMVERT(1099;2100;0.5;1))

7. H_o:"Die Testperson kann nicht mehrheitlich richtige Voraussagen machen".
H_A : Die Testperson kann mehrheitlich richtige Voraussagen machen ($\Longrightarrow$ rechtsseitiger Test).(Man könnte ähnlich wie oben auch für einen zweiseitigen Test argumentieren: wenn die Person systematisch falsch liegt, scheint sie hellseherische Fähigkeiten zu haben. Man könnte die Person in diesem Falle durchaus bei entsprechenden Spielen einsetzen, um oft zu gewinnen. Man müsste jeweils auf das Gegenteil dessen setzen, das sie voraussagt. In diesem Fall würden wir die Alternativhypothese wie folgt formulieren: Die Testperson kann mehrheitlich richtige oder mehrheitlich falsche Voraussagen machen). Als Signifikanzniveau legen wir 0.05 fest. Wir könnten z.B. die folgende Versuchsanordnung festlegen. Wir werfen 20 Münzen. Die entsprechenden Ergebnisse (= Summen der richtig Vorausgesagten) sind binomialverteilt ($B(20, 0.5)$). Wir testen rechtseitig. Mit Hilfe von Excel erhalten wir (s. Tabelle 3):

x_i	$P(X = x_i)$	$P(X \geq x_i)$
20	9.53674E-07	9.53674E-07
19	1.90735E-05	2.00272E-05
18	0.000181198	0.000201225
17	0.001087189	0.001288414
16	0.004620552	0.005908966
15	0.014785767	0.020694733
14	0.036964417	0.057659149
13	0.073928833	0.131587982

Tabelle 3: Berechnung des $0.05-$Verwerfungsbereich bei einer binomialverteilten Statistik; $X \sim B(20, 0.5)$); Verwerfungsbereich $= \{15, 16, 17, 18, 19, 20\}$

Beim festgelegten Signifikanzniveau wird die Nullhypothese verworfen, wenn die Testperson mindestens 15 richtige Voraussagen macht (da $P(X \geq 15) = 0.020694733 < 0.05$).

8. H_o : Die zwei Qualitäten können nicht auseinandergehalten werden.
H_A : Die zwei Qualitäten können auseinandergehalten werden, d.h. überwiegend richtig oder überwiegend falsch klassifiziert werden ($\Longrightarrow$ zweiseitiger Test).
Man wählt eine spezifische Anzahl m von Stichproben aus der ersten Qualität, und die restlichen $100 - m$ aus der anderen Qualität. Wir ordnen den Stichproben Zufallszahlen zu und ordnen sie der Grösse nach. Je zehn werden dann den Versuchspersonen zugeteilt. Wir können das ganze als einen Binomial-Test mit $X \sim B(100, 0.5)$ ansehen. Mit Excel erhalten wir: mindestens 64 müssen richtig klassiert werden oder höchstens $100 - 64 = 36$ (denn =BINOM.VERT(36;100;0.5;1)=0.00331856; $2 \cdot 0.00331856 = 0.006637\,1 < 0.01$, während

=BINOM.VERT(37;100;0.5;1) = 0.006016488 und $2 \cdot 0.006016488 = 0.012033 \geq 0.01$). (s. Tabelle 4)

x_i	$P(X = x_i)$	$P(X \geq x_i)$
66	0.000458105	0.000894965
65	0.000863856	0.001758821
64	0.001559739	0.00331856
63	0.002697928	0.006016488
62	0.00447288	0.010489368
61	0.007110732	0.0176001
60	0.010843867	0.028443967
59	0.015869073	0.04431304
		$P(X \leq x_i)$
34		0.000894965
35		0.001758821
36		0.00331856
37		0.006016488
38		0.010489368
39		0.0176001
40		0.028443967
41		0.04431304

Tabelle 4: Berechnung des 0.01−Verwerfungsbereich bei einer binomialverteilten Statistik; $X \sim B(100, 0.5)$); $P(X \leq 36) + P(X \geq 64) = 2 \cdot 0.00331856 = 0.006637121 < 0.01$; $P(X \leq 37) + P(X \geq 63) = 2 \cdot 0.006016488 = 0.012032976 > 0.01$; Verwerfungsbereich $= \{0, ..., 36, 64, ..., 100\}$

Wir könnten auch die Normalverteilung verwenden: $\frac{x-(100 \cdot 0.5)}{\sqrt{100 \cdot 0.5^2}} = 0.2x - 10 \sim N(0, 1)$.
$\Phi(0.2x - 10) = 0.995$, d.h. $\Phi^{-1}(0.995) = 0.2x - 10$. Wir bestimmen mit Hilfe von Excel $\Phi^{-1}(0.995)$ (mit norminv): 2.575834515. Wir erhalten somit: $0.2x - 10 = 2.575834515$, $\Longrightarrow$ $x = 62.879$. Wir müssten somit mindestens 63 richtige oder höchstens $100 - 63 = 37$ richtige Klassierungen haben - es liegt also eine kleine Abweichung vom exakten Resultat vor.

9. H_0 : Es hat nicht mehr Straftaten als vorher.
H_A : Es hat mehr Straftaten als vorher (rechtsseitiger Test).
Verteilung der Teststatistik: $X \sim B(12500 + 13000; 0.5)$. Den p-Wert können wir hier mit der Näherungsformel berechnen. $p-$Wert: $P(X \geq 13000) \approx 1 - \Phi\left(\frac{13000-np}{\sqrt{np(1-p)}}\right) = 1 - \Phi\left(\frac{13000-12750}{79.84359711}\right) = 1 - \Phi(3.131121455)$
$= 0.000870768 < 0.05$. Die Entwicklung lässt sich schlecht durch Zufall erklären.
Exakt: $P(X \geq 13000) = 1-$binom.vert$(12999, 12500 + 13000, 0.5, 1) = 0.00088918$

10. Bei jeder der 40 Aufgaben kann der Begutachtete zwischen falsch und wahr unterscheiden. H_0 : Der Begutachtete kann nicht zwischen wahr und falsch unterscheiden. H_A : Der Begutachtete kann zwischen wahr und falsch unterscheiden, wobei er oft das falsche wählen wird, weil er seine mathematische Unfähigkeit beweisen will (linksseitiger Test; er wählt weniger korrekte als jemand der zwischen wahr und falsch nicht unterscheiden kann). Wir berechnen $P(X \leq 11)$ für $X \sim B(40, 0.5)$. Wir erhalten: $P(X \leq 11) = \sum_{i=0}^{11} \binom{40}{i} 0.5^i (1 - 0.5)^{40-i} = 3.213\,3 \times 10^{-3} < 0.05$. Wir können die Nullhypothese verwerfen. Der Begutachtete kann zwischen falsch und wahr unterscheiden.

11. Die Teststatistik ist binomialverteilt mit $n = 10$ und $p = \ 0.1$. Damit gilt $P(X \geq 3) = 1 - P(X \leq 2) = 0.070190826$. Auf dem 0.05-Signifikanzniveau ist das Ergebnis fast signifikant. $P(X \geq 8) = 1 - P(X \leq 7) = 3.736 \cdot 10^{-7}$. Es handelt sich um das implizite Signifikanzniveau, das dem Test des Schweizerfernsehens zu Grunde liegt. Es ist wohl unvernünftig tief angesetzt.

12. siehe als Beispiele früherer Jahrgänge auf der Home-Page http://math.logik.ch „Beispiele_diskrete_Tests.pdf“.

8 Lernziele

- Grundbegriffe der Testtheorie kennen und korrekt verwenden können (Nullhypothese, Alternativhypothese, Signifikanzniveau, rechts- links- und beidseitiger Test, p-Wert, Fehler 1. und 2. Art, Teststatistik, Testwert, Verwerfungsbereich).
- Verwerfungsprinzip anwenden können.
- Tests mit diskreten Verteilungen durchführen können (insbesondere Binomial- und Poissonverteilung, sowie hypergeometrische Verteilung).
- Übungen von der Art des Übungsblocks lösen können.

9 Bemerkung

Verfasst von Paul Ruppen, Statistiksupport der PH Wallis, 20. September 2021