2 Statistische Tests

Zu den grundlegenden Aufgaben der mathematischen Statistik zählt neben dem Schätzen von Parametern auch das Testen von Hypothesen, d.h. gewisser Annahmen über Grundgesamtheiten. Hypothesen können sich z.B. auf frühere Erfahrungswerte stützen, sie können einen Sollwert darstellen oder das Ergebnis einer zu verifizierenden Theorie sein. Mit einem statistischen Test prüft man also eine vorgefasste Vermutung über eine oder auch mehrere Grundgesamtheiten anhand von Zufallsstichproben. Zunächst werden wir uns mit Tests über Parameter von Verteilungen befassen, deren Typ (zumeist die Normalverteilung) bekannt ist. Man spricht in diesem Zusammenhang von **parametrischen** oder auch **verteilungsabhängigen Verfahren**. Im darauf folgenden Abschnitt wird ein Überblick über Testverfahren gegeben, mit deren Hilfe man den Verteilungstyp selbst überprüfen kann, sowie Verfahren, die gänzlich unabhängig von der Verteilung der betrachteten Grundgesamtheit sind. Man nennt daher solche Methoden **nichtparametrische** oder auch **verteilungsunabhängige Verfahren**.

2.1 Parametrische Testverfahren

Das erste Beispiel eines statistischen Testverfahrens, mit dem wir uns in diesem Abschnitt beschäftigen, lautet wie folgt:

(i) Ein einführendes Beispiel: Überprüfung einer Hypothese über den Mittelwert μ einer normalverteilten Grundgesamtheit mit bekannter Varianz σ^2.

Beispiel: Wenden wir uns nochmals der Produktion elektrischer Bauteile und der Frage nach der Größe eines Bauteils zu. Wir nehmen an, dass die betreffende Größe durch eine normalverteilte Zufallsvariable X beschrieben wird, wobei deren Varianz $\sigma^2 = 24^2$ aus früheren Messungen bekannt sei. Eine konkrete Stichprobe von n = 30 Messwerten liefere den Mittelwert $\bar{x} = 290{,}5$. Kann dann aufgrund dieser Stichprobe angenommen werden, dass der Erwartungswert μ von X dem Sollwert $\mu_0 = 300$ entspricht?

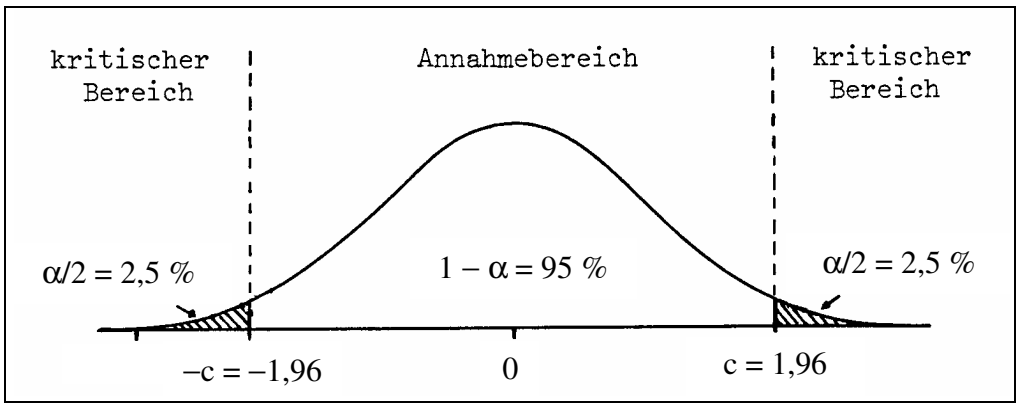
Zur Beantwortung dieser Frage formulieren wir zunächst die so genannte Nullhypothese H_0: $\mu = \mu_0$ bzw. deren Negation, die Alternative (Alternativhypothese) H_1: $\mu \neq \mu_0$. Unsere Aufgabe besteht nun darin, festzustellen, ob die Nullhypothese anhand der gegebenen Stichprobe beibehalten werden kann oder aber abzulehnen (und damit die Alternativhypothese anzunehmen) ist. Diesen Entscheidungsprozeß kann man folgendermaßen führen:

Ausgangspunkt unserer Überlegungen ist eine normalverteilte Grundgesamtheit X vom Typ $N(\mu,\sigma^2)$. Dann ist das Stichprobenmittel $\overline{X}$ – wie wir wissen – ebenfalls normalverteilt mit dem Erwartungswert $E(\overline{X}) = \mu$ und der Varianz $Var(\overline{X}) = \sigma^2 / n$. Dies bedeutet, dass bei Bestehen der Nullhypothese H_0: $\mu = \mu_0$ die standardisierte Zufallsvariable

$$TG = \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n}$$

standardnormalverteilt ist. Die Größe TG wird als **Testgröße** oder **Teststatistik** bezeichnet.

Man kann die Nullhypothese sicher nicht bereits dann ablehnen, wenn für eine konkrete Stichprobe $\bar{x} < \mu_0$ oder $\bar{x} > \mu_0$ ist, denn die Wahrscheinlichkeit dafür ist jeweils 1/2. Vernünftig erscheint es aber, die Nullhypothese abzulehnen, wenn $\bar{x}$ den Sollwert μ_0 „wesentlich“ unter- oder überschreitet, d.h., wenn der mit Hilfe von $\bar{x}$ errechnete Wert von TG gewisse kritische Grenzen $-c$ bzw. $+c$ unter- bzw. überschreitet, die so gewählt sind, dass eine Unterschreitung (und ebenso auch eine Überschreitung) nur mit einer vorgegebenen kleinen Wahrscheinlichkeit $\alpha/2$ zu erwarten ist. Für den **kritischen Wert** c erhält man aus der Forderung $P(TG < -c) = P(TG > c) = \alpha/2$ die Bestimmungsgleichung $\Phi(c) = 1 - \alpha/2$.



Fällt die aus den Stichprobenwerten $x_1, x_2, ..., x_n$ berechnete Realisierung von TG in den so genannten **Annahmebereich**, ist also $-c \le TG \le c$, so wird die Nullhypothese beibehalten, fällt sie hingegen in den **kritischen Bereich**, d.h. gilt $TG < -c$ oder $TG > c$, so wird die Nullhypothese verworfen. (Wir werden im folgenden zwischen der Zufallsvariablen TG und ihrer Realisierung nicht mehr unterscheiden, sondern in beiden Fällen einfach von der Testgröße TG sprechen.) Dabei ist c durch die vorzugebende Wahrscheinlichkeit α (meist 5%, 1% oder 0.1%) bestimmt. Aus der Tafel für die Verteilungsfunktion der Standardnormalverteilung (im Anhang) entnehmen wir beispielsweise für $\alpha = 0{,}05$ und $\Phi(c) = 1 - \alpha/2 = 0{,}975$ den Wert $c = 1{,}96$. Die Zahl α heißt **Signifikanzniveau** oder auch **Irrtumswahrscheinlichkeit**, denn sie gibt die Wahrscheinlichkeit dafür an, dass die Nullhypothese abgelehnt wird, obwohl sie richtig ist (vgl. Abbildung).

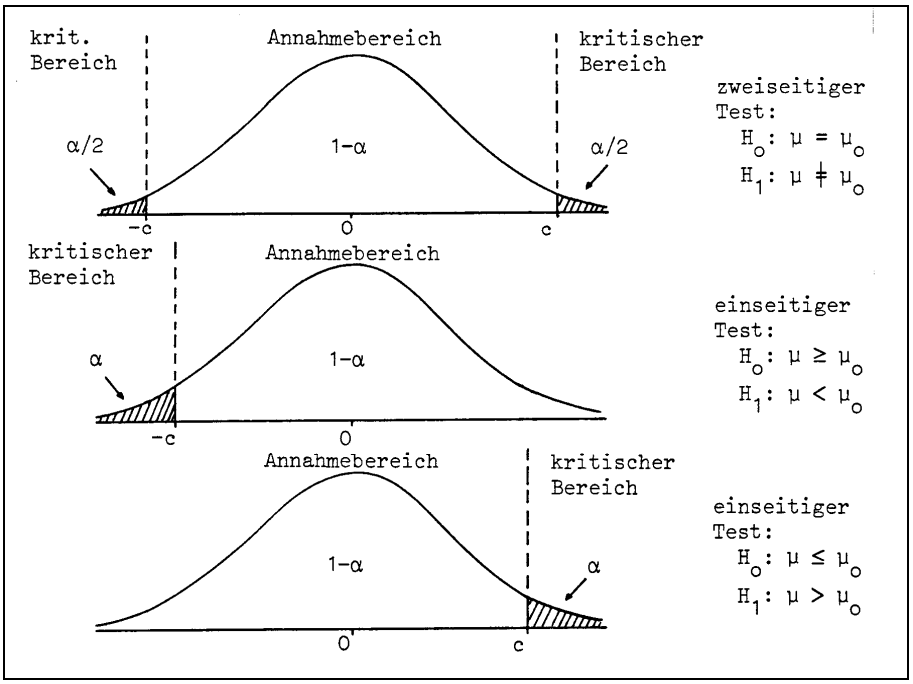
Beispiel (Fortsetzung): Wir kommen nun auf obiges Beispiel zurück und testen die Hypothese H_0: $\mu = 300$, dass die Bauteilgröße dem Sollwert $\mu_0 = 300$ entspricht, und zwar auf dem Signifikanzniveau $\alpha = 0{,}05$. Der zugehörende kritische Wert beträgt $c = 1{,}96$. Wir berechnen nun aus dem Mittelwert $\bar{x} = 290{,}5$ die Testgröße

$$TG = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = \frac{290{,}5 - 300}{24}\sqrt{30} = -2{,}17\,.$$

Da $TG < -c$, fällt die Prüfgröße in den kritischen Bereich und die Nullhypothese H_0 ist auf dem 5%-Signifikanzniveau abzulehnen. Es ist also anzunehmen, dass sich die Produktion aufgrund irgendwelcher Störursachen vom Sollwert entfernt hat.

Hätten wir jedoch die Irrtumswahrscheinlichkeit $\alpha = 0{,}01$ gewählt, so ergibt sich gemäß $\Phi(c) = 1 - \alpha/2 = 0{,}995$ der kritische Wert $c = 2{,}58$. TG liegt also diesmal im Annahmebereich, d.h., die Nullhypothese H_0 kann auf dem 1%-Signifikanzniveau nicht mehr abgelehnt werden.

In der betrachteten Testsituation wird die Nullhypothese dann verworfen, wenn der Absolutbetrag der Testgröße TG eine bestimmte kritische Schranke überschreitet, d.h., wenn $\bar{X}$ in einer Stichprobe signifikant unter oder über dem festen Wert μ_0 liegt. Man spricht daher von einem **zweiseitigen Test**. Abweichungen können aber auch nur in einer Richtung bedeutungsvoll oder möglich sein. Man denke z.B. an die Überprüfung einer Maximalkonzentration im Gesundheits- oder Umweltbereich. In diesem Fall lautet die Nullhypothese H_0: $\mu \le \mu_0$, die Alternative H_1: $\mu > \mu_0$. Man wird also die Nullhypothese nur dann verwerfen, wenn eine signifikante Abweichung des Stichprobenmittels nach oben auftritt, eine Abweichung nach unten spielt keine Rolle. Der kritische Wert c wird dann aus der Gleichung $\Phi(c) = 1 - \alpha$ bestimmt, wobei Φ wieder die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Gilt für den Wert TG der Teststatistik TG > c, so wird die Hypothese H_0 verworfen, in jedem anderen Fall wird sie beibehalten. Wir nennen den so modifizierten Test einen **einseitigen Test auf Überschreitung**. Analog ist ein **einseitiger Test auf Unterschreitung** möglich. Annahmebereiche und kritische Bereiche beim ein- und zweiseitigen Test sind in folgender Abbildung dargestellt.



Beispiel (Fortsetzung): Wäre in obigem Beispiel eine Stichprobe mit dem Mittelwert $\bar{x} = 305$ bereits eine signifikante Überschreitung des Sollwerts $\mu_0 = 300$?

Diese Fragestellung führt auf einen Test auf Überschreitung. Die Nullhypothese lautet in diesem Fall H_0: $\mu \le 300$, die Alternative H_1: $\mu > 300$. Zum Signifikanzniveau $\alpha = 0{,}05$ bestimmt man den kritischen Wert c aus der Gleichung $\Phi(c) = 1 - \alpha = 0{,}95$, also $c = 1{,}65$. Schließlich berechnet man für die Testgröße

$$TG = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = \frac{305 - 300}{24}\sqrt{30} = 1{,}14 < c\,.$$

Also liegt TG im Annahmebereich und die Nullhypothese wird auf dem 5%-Signifikanzniveau beibehalten.

Allgemein läuft jeder Signifikanztest nach dem folgenden Schema ab:

- Formulierung der Nullhypothese H_0 und der Alternative H_1
- Bereitstellung einer geeigneten Testgröße TG
- Wahl des Signifikanzniveaus α und Bestimmung des kritischen Bereichs für die vorgegebene Testgröße
- Berechnung der Testgröße TG aus einer Zufallsstichprobe
- Ablehnung der Nullhypothese, falls TG in den kritischen Bereich fällt

Fällt der Wert der Teststatistik in den Annahmebereich, so kann diese Tatsache noch nicht als eine Bestätigung der Nullhypothese gewertet werden. Ein Wert im Annahmebereich besagt lediglich, dass die Hypothese zu den Daten der Stichprobe nicht im Widerspruch steht. Da also das Verwerfen einer Hypothese mehr Aussagekraft besitzt als deren Beibehaltung, ist es günstig, nach Möglichkeit die Nullhypothese als Gegenteil von dem zu formulieren, was man beweisen möchte, und zu versuchen, dieses Gegenteil zu widerlegen.

Im Zusammenhang mit diesen Überlegungen muss betont werden, dass das Signifikanzniveau prinzipiell vor der Durchführung des Testverfahrens festgesetzt werden sollte. Trotzdem hat es sich in der Praxis – insbesondere beim Arbeiten mit einschlägigen Computerprogrammen – eingebürgert, oft erst nach dem Test das **erreichte Signifikanzniveau** P zu bestimmen und in der Form $P > 0{,}05$ (nicht signifikant), $P \leq 0{,}05$, $P \leq 0{,}01$ bzw. $P \leq 0{,}001$ (signifikant) anzugeben. Dieser so genannte **P-Wert** entspricht der Wahrscheinlichkeit für eine Abweichung der Testgröße vom Sollwert im beobachteten oder noch größeren Ausmaß (im Sinne der Alternativhypothese). Es bleibt dann dem Anwender überlassen, das Testergebnis gemäß dem von ihm bevorzugten oder für dieses Problem als geeignet erachteten Signifikanzniveau zu bewerten. In diesem Fall ist dann der letzte Punkt in oben angegebenem Schema zu ersetzen durch

- Ablehnung der Nullhypothese, falls P-Wert kleiner als α

Zum Abschluss der einführenden Bemerkungen zur Testtheorie wollen wir uns noch mit den möglichen Fehlentscheidungen befassen, die beim Testen von Hypothesen auftreten können. Wie wir gesehen haben, wird durch die Signifikanzzahl α der kritische Bereich festgelegt, also die Menge jener Werte der Teststatistik, die zu einem Verwerfen der Nullhypothese führen. Nun kann es natürlich passieren, dass die Stichprobe einen so „unwahrscheinlichen" Wert liefert, dass die Nullhypothese abgelehnt wird, obwohl sie richtig ist. Dabei begeht man einen Irrtum; dieser wird **Fehler erster Art** oder **α-Fehler** genannt, da er mit der Wahrscheinlichkeit α auftritt. Das Signifikanzniveau α gibt also die Wahrscheinlichkeit an,

die Nullhypothese zu unrecht abzulehnen. Einen Fehler begeht man aber auch dann, wenn man die Nullhypothese nicht verwirft, obwohl sie in Wirklichkeit falsch ist. Diese Fehlentscheidung wird **Fehler zweiter Art** oder **β-Fehler** genannt, wo β gerade die Wahrscheinlichkeit für sein Auftreten bezeichnet. In wirtschaftlichem Zusammenhang spricht man oft auch vom **Produzentenrisiko** (für den Fehler 1. Art) und vom **Konsumentenrisiko** (für den Fehler 2. Art).

Mögliche Entscheidungen und deren Wahrscheinlichkeiten beim Hypothesentest:		
Die Nullhypothese	**wird angenommen**	**wird verworfen**
ist richtig	richtige Entscheidung $1 - \alpha$	Fehler 1. Art α
ist falsch	Fehler 2. Art β	richtige Entscheidung $1 - \beta$

Genau genommen hängt die Wahrscheinlichkeit, dass die Nullhypothese abgelehnt wird, vom aktuellen (wenn auch unbekannten) Parameter der Verteilung der Grundgesamtheit ab, im konkreten Fall also vom Mittelwert μ. Diese Wahrscheinlichkeit bezeichnet man als **Gütefunktion** oder **Trennschärfe** g(μ) des Tests. Liegt μ im Bereich der Nullhypothese, so gilt stets $g(\mu) \leq \alpha$, liegt μ hingegen im Bereich der Alternative, dann ist $g(\mu) = 1 - \beta$. Mit Hilfe der Gütefunktion ist ein Vergleich mehrerer Testverfahren zur selben Nullhypothese und zum selben Signifikanzniveau möglich: Je größer die Güte eines Tests für Parameterwerte im Bereich der Alternative ist, desto geringer ist die Wahrscheinlichkeit, einen Fehler 2. Art zu begehen, d.h., die Nullhypothese zu Unrecht anzunehmen. Je größer also die Güte eines Tests ist, desto besser kann der Test einen bestehenden Unterschied erkennen. Ganz wesentlich wird die Güte vom Umfang der dem Test zugrunde liegenden Stichprobe beeinflusst: Mit wachsendem Umfang steigt auch die Güte.

Im Allgemeinen wird man natürlich bestrebt sein, beide Fehlerwahrscheinlichkeiten α und β so gering wie möglich zu halten, wobei allerdings zu beachten ist, dass z.B. eine Verringerung von α meist eine Zunahme von β zur Folge hat und umgekehrt. Man wird also bei der Wahl des Signifikanzniveaus im Auge behalten müssen, welche Konsequenzen mit einem Fehler erster bzw. zweiter Art verbunden sind. Es gilt die folgende Faustregel: Sind die Konsequenzen schwerwiegend, wenn die Nullhypothese falsch ist, aber irrtümlich nicht verworfen wird, dann wird man die Irrtumswahrscheinlichkeit α eher groß wählen. Überwiegen dagegen die Nachteile, wenn H_0 richtig ist, jedoch die Alternative irrtümlich angenommen wird, so ist eine kleine Irrtumswahrscheinlichkeit α vorzuziehen.

Vergleich von Mittelwerten

In diesem Abschnitt werden Testverfahren zum Vergleich von Mittelwerten für ein und zwei Stichproben behandelt. Der Test einer Hypothese $\mu = \mu_0$ über den Mittelwert μ einer normalverteilten Grundgesamtheit mit bekannter Varianz wurde bereits im letzten Abschnitt ausführlich behandelt. In der Regel wird aber die Varianz σ^2 unbekannt sein und muss durch die Stichprobenvarianz s^2 geschätzt werden. In diesem Fall kommt ein modifiziertes Testverfahren, der so genannte **Einstichproben-t-Test** zur Anwendung.

(ii) Test des Mittelwerts μ einer Normalverteilung mit unbekannter Varianz σ^2: Einstichproben-t-Test

Zur Überprüfung der Hypothese H_0: $\mu = \mu_0$ über den Mittelwert einer Normalverteilung geht man wie beim Test (i) vor, wobei lediglich die unbekannte Varianz σ^2 durch die Stichprobenvarianz s^2 ersetzt wird. Die Rolle der Teststatistik übernimmt also die aus den Stichprobenvariablen $\overline{X}$ und S^2 gebildete Stichprobenfunktion

$$TG = \frac{\overline{X} - \mu_0}{S}\sqrt{n}\,.$$

Sie besitzt unter Annahme der Nullhypothese H_0 eine t-Verteilung mit $n - 1$ Freiheitsgraden, wobei n den Umfang der Stichprobe bezeichnet.

Das Entscheidungsverfahren beim Einstichproben-t-Test, das sowohl als zweiseitiger wie als einseitiger Test durchgeführt werden kann, läuft nach dem gewohnten Schema ab: Zur Signifikanzzahl α bestimmt man den kritischen Wert c aus der Gleichung $F(c) = 1 - \alpha/2$ (bzw. $F(c) = 1 - \alpha$ beim einseitigen Test) für die Verteilungsfunktion F der t-Verteilung mit $FG = n - 1$ Freiheitsgraden. Dann berechnet man anhand der gegebenen Stichprobe die Werte $\bar{x}$, s^2 und schließlich den Wert der Prüfgröße TG. Fällt TG in den kritischen Bereich, d.h. gilt $|TG| > c$ (bzw. $TG < -c$ oder $TG > c$ beim einseitigen Test auf Unter- oder Überschreitung), so wird die Nullhypothese verworfen.

Beispiel: Die Gewichtsverteilung von Neugeborenen werde durch eine Normalverteilung mit den Parametern μ und σ^2 beschrieben. In einer Stichprobe von 47 Neugeborenen wurde ein Mittelwert von 3281 g und eine Standardabweichung von 705 g festgestellt. Kann aufgrund dieser Daten geschlossen werden, dass das mittlere Geburtsgewicht $\mu = 3200$ g beträgt?

Wir formulieren die zweiseitige Nullhypothese H_0: $\mu = \mu_0$ mit $\mu_0 = 3200$ und die Alternative H_1: $\mu \neq \mu_0$. Als Signifikanzniveau wählen wir $\alpha = 0{,}05$ und bestimmen aus $F(c) = 1 - \alpha/2 = 0{,}975$ für die t-Verteilung mit $FG = 46$ den kritischen Wert $c = 2{,}01$. Als Realisierung der Testgröße TG erhalten wir dann

$$TG = \frac{3281 - 3200}{705}\sqrt{47} = 0{,}79\,.$$

Wegen $|TG| \leq c$ müssen wir die Nullhypothese beibehalten. Die vorliegenden Daten sprechen also nicht für ein von der Marke 3200 g abweichendes durchschnittliches Geburtsgewicht.

Wir haben bisher den Mittelwert einer Normalverteilung mit einer gegebenen Zahl verglichen und wollen nun die Mittelwerte zweier Grundgesamtheiten zueinander in Beziehung setzen. Von jeder der beiden Grundgesamtheiten X_1 bzw. X_2 liege eine Stichprobe

$$x_{11}, x_{12}, ..., x_{1n_1} \text{ bzw. } x_{21}, x_{22}, ..., x_{2n_2}$$

vor. Diese Stichproben können entweder **abhängig** oder **unabhängig** sein. Abhängige (oder verbundene) Stichproben zeichnen sich dadurch aus, dass sämtliche Stichprobenwerte in

Paaren vorliegen (z.B. weil sie an derselben Untersuchungseinheit, etwa an derselben Person gemessen wurden). Folglich stimmen dann auch die beiden Stichprobenumfänge n_1 und n_2 überein. Zwei Stichproben heißen hingegen unabhängig, wenn ihre Beobachtungswerte nicht unmittelbar in einen paarweisen Zusammenhang gebracht werden können.

(iii) Vergleich der Mittelwerte μ_1 und μ_2 zweier Normalverteilungen unter Benutzung verbundener Stichproben: Differenzen-t-Test

Zum Test der Hypothese H_0: $\mu_1 = \mu_2$ für die Mittelwerte zweier Normalverteilungen unter Benützung verbundener Stichproben (**Differenzen-t-Test**) geht man wie folgt vor: Man bildet zunächst die Differenzen entsprechender Stichprobenwerte. Für die aus diesen Differenzen gebildete Stichprobe prüft man dann mit dem Einstichproben-t-Test, ob sie aus einer Verteilung mit dem Mittelwert $\mu = 0$ stammt.

(iv) Vergleich der Mittelwerte μ_1 und μ_2 zweier Normalverteilungen unter Benutzung unabhängiger Stichproben: Zweistichproben-t-Test

Zwei normalverteilte Grundgesamtheiten bilden auch das statistische Modell für den Vergleich der Mittelwerte μ_1 und μ_2 zweier Normalverteilungen unter Benützung unabhängiger Stichproben (**Zweistichproben-t-Test**). Die zugehörenden Varianzen brauchen nicht bekannt zu sein, sie werden aber als gleich vorausgesetzt. Bezeichnen n_1, $\overline{X}_1$ bzw. S_1^2 Stichprobenumfang, Mittelwert bzw. Varianz der ersten Stichprobe und n_2, $\overline{X}_2$ bzw. S_2^2 die entsprechenden Größen der zweiten Stichprobe, so bildet man die Zufallsvariable

$$TG = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(n_1-1)S_1^2 + (n_2-1)S_2^2}} \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}} .$$

Diese besitzt unter der Annahme $\mu_1 = \mu_2$ wieder eine t-Verteilung, und zwar mit $n_1 + n_2 - 2$ Freiheitsgraden. Damit ist auch der Testablauf vorgezeichnet:

Im Fall eines zweiseitigen Tests mit der Nullhypothese H_0: $\mu_1 = \mu_2$ bestimmt man zur Signifikanzzahl α den kritischen Wert c aus $F(c) = 1 - \alpha/2$ für die t-Verteilung mit $FG = n_1 + n_2 - 2$. Dann berechnet man für beide Stichproben Mittelwerte und Varianzen und setzt diese schließlich in die Formel für die Testgröße TG ein. Gilt für den so erhaltenen Wert TG der Teststatistik $|TG| > c$, so wird die Nullhypothese verworfen; ist $|TG| \leq c$, so wird sie beibehalten. Beim einseitigen Test ist entsprechend vorzugehen.

Beispiel: Das Gewicht von Neugeborenen sei normalverteilt mit den Parametern μ_1 und σ^2 für Knaben bzw. μ_2 und σ^2 für Mädchen. Man teste die Hypothese $\mu_1 = \mu_2$ auf dem 5%-Signifikanzniveau anhand einer Stichprobe von $n_1 = 29$ männlichen sowie einer Stichprobe von $n_2 = 18$ weiblichen Neugeborenen, wobei die empirischen Werte $\overline{x}_1 = 3170$ g, $s_1 = 637$ g, $\overline{x}_2 = 3460$ g und $s_2 = 789$ g ermittelt wurden.

Wir führen einen zweiseitigen Test für die Nullhypothese H_0: $\mu_1 = \mu_2$ mit der Alternative H_1: $\mu_1 \neq \mu_2$ unter Benützung von zwei unabhängigen Stichproben durch. Zur Signifikanzzahl $\alpha = 0{,}05$ bestimmen wir zunächst aus $F(c) = 1 - \alpha/2 = 0{,}975$ für die t-Verteilung mit FG = 45 den kritischen Wert c = 2,01. Die Prüfgröße TG nimmt für die gegebenen Stichproben den Wert

$$TG = \frac{3170 - 3460}{\sqrt{28 \cdot 637^2 + 17 \cdot 789^2}} \sqrt{\frac{29 \cdot 18 \cdot 45}{47}} = -1.38$$

an, d.h., sie liegt deutlich im Annahmebereich. Somit kann aufgrund der vorliegenden Daten kein signifikanter Unterschied zwischen den Gewichten neugeborener Knaben bzw. Mädchen festgestellt werden.

(v) Test für die Wahrscheinlichkeit p = P(A) eines Ereignisses A: Binomialtest

Ebenfalls auf einen Mittelwertvergleich zurückgeführt werden kann der Test für die Wahrscheinlichkeit p eines Ereignisses A, d.h. der Test für den Parameter p einer Binomialverteilung. Wir beschreiben – wie bereits im vorhergehenden Kapitel – die n-malige Ausführung des zum Ereignis A gehörenden Zufallsexperiments durch die Zufallsvariablen $X_1, ..., X_n$, welche nur die Werte 1 (mit Wahrscheinlichkeit p) oder 0 (mit Wahrscheinlichkeit 1 − p) annehmen können. Dann ist die Zufallsvariable $H = (X_1 + ... + X_n)/n$, also die relative Häufigkeit des Ereignisses A in einer Stichprobe vom Umfang n, annähernd normalverteilt nach $N(p, p(1-p)/n)$. Unter der Annahme $p = p_0$ ist daher die Testgröße

$$TG = \frac{H - p_0}{\sqrt{p_0 (1 - p_0)}} \sqrt{n}$$

approximativ standardnormalverteilt (falls $np(1-p) \geq 9$). Mittels dieser Teststatistik kann nun ein Test der Nullhypothese H_0: $p = p_0$ für die Wahrscheinlichkeit p des Ereignisses A nach dem üblichen Muster durchgeführt werden.

Beispiel: Mit einem gewöhnlichen Würfel wird unter n = 100 Würfen genau 20 mal ein Sechser gewürfelt. Kann der Würfel noch als fair angesehen werden?

Wir testen die Hypothese p = 1/6 auf dem 5%-Signifikanzniveau. Der kritische Wert c zu $\alpha = 0{,}05$ wird diesmal wieder aus der Standardnormalverteilung gemäß $\Phi(c) = 1 - \alpha/2 = 0{,}975$ bestimmt und ergibt sich zu c = 1,96. Mit n = 100, h = 20/100 = 1/5 für die relative Häufigkeit und $p_0 = 1/6$ berechnen wir

$$TG = \frac{\frac{1}{5} - \frac{1}{6}}{\sqrt{\frac{1}{6}(1 - \frac{1}{6})}} \sqrt{100} = 0{,}89 .$$

Wegen |TG| < c wird die Nullhypothese beibehalten, also ist der Würfel fair.

Beispiel: Bei einer Umfrage zur zahnmedizinischen Vorsorge gaben 156 von 360 befragten Personen an, regelmäßig zweimal pro Jahr einen Zahnarzt aufzusuchen. Lässt diese Umfrage die Vermutung zu, dass der entsprechende Anteil in der Bevölkerung unter 50 % liegt?

Wir führen einen einseitigen Test auf Unterschreitung durch, d.h., wir testen die Hypothese H_0: $p \geq 0{,}5$ gegen die Alternative H_1: $p < 0{,}5$. Zur Signifikanzzahl $\alpha = 0{,}05$ bestimmen wir nun den kritischen Wert c aus der Gleichung $\Phi(c) = 1 - \alpha = 0{,}95$, also $c = 1{,}65$. Die Berechnung der Testgröße TG liefert schließlich den Wert

$$TG = \frac{\frac{156}{360} - 0{,}5}{\sqrt{0{,}5(1-0{,}5)}}\sqrt{360} = -2{,}53 .$$

Wegen $TG < -c$ liegt die Teststatistik im kritischen Bereich und die Nullhypothese muss verworfen werden, die obige Vermutung wird bestätigt.

Weitere parametrische Testverfahren

Analog zu den bisher beschriebenen Mittelwertvergleichen kann – auf Basis einer normalverteilten Grundgesamtheit – mit geeigneten Tests überprüft werden, ob die Varianz σ^2 der Grundgesamtheit mit einem vorgegebenen Wert σ_0^2 übereinstimmt (χ^2-Streuungstest) oder ob die Varianzen σ_1^2 und σ_2^2 zweier Grundgesamtheiten gleich sind (F-Test).

Ein sehr allgemeines und in vielen Fällen außerordentlich leistungsfähiges statistisches Verfahren stellt die Methode der **Varianzanalyse (ANOVA)** dar. Die Varianzanalyse ist im einfachsten Fall eine Verallgemeinerung des Zweistichproben-t-Tests und erlaubt einen Mittelwertvergleich für mehrere normalverteilte Grundgesamtheiten, also z.B. den Vergleich mehrerer Produktionsverfahren oder Behandlungsmethoden. Auf diese Weise kann der Einfluss eines oder auch mehrerer qualitativer Merkmale (genannt Faktoren) auf ein metrisches Merkmal untersucht werden. Varianzanalytische Verfahren sind i. Allg. sehr rechenaufwendig und erfordern den Einsatz geeigneter Statistikprogramme.

2.2 Nichtparametrische Testverfahren

In vielen Fällen wird die bisher gemachte Annahme über normalverteilte Grundgesamtheiten nicht erfüllt sein, so dass nichtparametrische bzw. verteilungsunabhängige Verfahren zur Verwendung kommen. Dazu zählen einerseits Schätz- und Testverfahren für Parameter von nicht normalverteilten Grundgesamtheiten, anderseits auch Tests über die Gestalt einer Verteilung überhaupt (Anpassungstests).

Der Vorteil nichtparametrischer Methoden liegt also zunächst darin, dass sie nur sehr schwache Annahmen über die Grundgesamtheit voraussetzen. Darüber hinaus sind sie nicht nur für metrische Skalen, sondern z.B. als Rangtests auch für ordinales Messniveau geeignet. Der Nachteil dieser Verfahren liegt in ihrer geringeren Güte, d.h., ein bestehender Lageunterschied etwa wird im Vergleich zu parametrischen Tests weniger oft erkannt. Zum Vergleich zweier Testverfahren definiert man die **Effizienz** oder **Wirksamkeit** eines Tests im Vergleich zu einem anderen als Verhältnis der Stichprobenumfänge, die zur Erreichung der gleichen Güte notwendig wären. Also lässt sich für einen nichtparametrischen Test die gleiche Güte erzielen wie bei einem entsprechenden Parametertest, wenn nur der Stichprobenumfang der Effizienz entsprechend vergrößert wird.

Nichtparametrische Tests für Lageparameter

Zum Lagevergleich für normalverteilte Grundgesamtheiten haben wir bisher die verschiedenen Varianten des t-Tests kennengelernt. Im Folgenden soll nun die Einschränkung auf normalverteilte Grundgesamtheiten aufgehoben werden, und wir setzen lediglich voraus, dass die betrachteten Grundgesamtheiten eine stetige Verteilung besitzen. In diesem Fall erweist sich der Median der Verteilung, und nicht wie bisher der Mittelwert, als geeigneter Lageparameter. (Bei symmetrischen Verteilungen fallen Mittelwert und Median ohnehin zusammen, so dass eine Unterscheidung in der Praxis vielfach belanglos ist.) Als nichtparametrisches Gegenstück zum Ein- bzw. Zweistichproben-t-Test werden wir nun den Vorzeichentest bzw. den U-Test besprechen.

(i) Vorzeichentest

Mit dem Vorzeichentest kann man prüfen, ob der Median einer Verteilung mit einem vorgegebenen Wert übereinstimmt, oder aber man vergleicht mittels verbundener Stichproben die Mediane verschiedener Grundgesamtheiten untereinander.

Beispiel: In einem Supermarkt soll die Wartezeit von Kunden an den Kassen untersucht werden. Von der Geschäftsleitung wird angestrebt, dass Kunden im Mittel höchstens drei Minuten an der Kasse warten müssen. Eine Stichprobe von 12 zufällig ausgesuchten Kunden ergibt folgende Werte (Wartezeit in Minuten):

5 4,5 3,5 6 1,5 12 15 3,5 5 7,5 2,5 4

Widerlegt diese Stichprobe die Vorgaben der Geschäftsleitung?

Falls für die mittlere Wartezeit, genauer gesagt für den Median $\tilde{\mu}$ der Wartezeiten $\tilde{\mu} = 3$ gilt, müsste etwa die Hälfte aller Wartezeiten kleiner als 3 und die Hälfte größer als 3 sein. Die Wahrscheinlichkeit, länger als 3 Minuten an einer Kasse warten zu müssen, wäre also für jeden Kunden genau p = 1/2. Somit beträgt die Wahrscheinlichkeit dafür, dass die Wartezeit für (mindestens) 10 von 12 Kunden – wie in der beobachteten Stichprobe – über 3 Minuten liegt, nach der Formel für die Binomialverteilung B(n = 12, p = 0,5) gerade

$$P(D^+ \geq 10) = \left(\binom{12}{10} + \binom{12}{11} + \binom{12}{12}\right)\left(\frac{1}{2}\right)^{12} = 0,0193,$$

wo D^+ die Anzahl der Werte bezeichnet, welche größer als 3 sind. Mit einem so unwahrscheinlichen Ergebnis brauchen wir, wenn das Signifikanzniveau etwa mit $\alpha = 5\%$ festgelegt ist, nicht zu rechnen. Damit ist die Vorgabe der Geschäftsleitung widerlegt.

Auf dieser Schlussweise beruht der **Vorzeichentest**. Man überprüft damit die Hypothese, dass die Merkmalsausprägungen einer Grundgesamtheit eine vorgegebene Zahl $\tilde{\mu}_0$ mit gleicher Wahrscheinlichkeit über- bzw. unterschreiten, d.h., ob der Median $\tilde{\mu}$ der Grundgesamtheit mit $\tilde{\mu}_0$ übereinstimmt. Die Nullhypothese für den zweiseitigen Vorzeichentest lautet daher H_0: $\tilde{\mu} = \tilde{\mu}_0$, die Alternative H_1: $\tilde{\mu} \neq \tilde{\mu}_0$. Die Entscheidung für eine der beiden Hypothesen

wird nun folgendermaßen anhand einer Zufallsstichprobe $X_1, \ldots, X_n$ durchgeführt. Gegebenenfalls werden alle Stichprobenwerte, die mit $\tilde{\mu}_0$ zusammenfallen, ausgeschieden und der Stichprobenumfang entsprechend reduziert. Dann ermittelt man die Testgröße

$TG = D^+ =$ Anzahl aller Stichprobenwerte, welche größer als $\tilde{\mu}_0$ sind.

Falls H_0 zutrifft, gilt $P(X_i < \tilde{\mu}_0) = P(X_i > \tilde{\mu}_0) = 1/2$ für alle i, und die Testgröße TG ist eine binomialverteilte Zufallsvariable mit den Parametern n und p = 1/2, d.h.

$$P(TG = k) = \binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k} = \binom{n}{k}\left(\frac{1}{2}\right)^n \quad \text{für } k = 0,\ldots,n.$$

Man bestimmt also zu vorgegebenem Signifikanzniveau α ganzzahlige kritische Werte c_1 und c_2 gemäß

$$P(TG < c_1) \le \alpha/2 \text{ und } P(TG > c_2) \le \alpha/2.$$

(Da die Größe TG eine diskrete Zufallsvariable darstellt, kann hier nur ≤ anstelle von = geschrieben werden, so dass das Testniveau α in der Regel nicht voll ausgeschöpft wird.) Die kritischen Werte c_1 und c_2 sind zu gegebenem Stichprobenumfang n und Signifikanzniveau α im Anhang tabelliert. Gilt nun für die konkrete Stichprobe, dass $TG < c_1$ oder $TG > c_2$, so wird die Nullhypothese H_0 abgelehnt, im Fall $c_1 \le TG \le c_2$ wird sie beibehalten.

Ganz analog geht man beim einseitigen Vorzeichentest vor. Beim Test auf Unterschreitung lautet die Nullhypothese H_0: $\tilde{\mu} \ge \tilde{\mu}_0$ gegenüber der Alternative H_1: $\tilde{\mu} < \tilde{\mu}_0$, beim Test auf Überschreitung wählt man H_0: $\mu \le \tilde{\mu}_0$ gegen H_1: $\tilde{\mu} > \tilde{\mu}_0$. Abgelehnt wird jeweils dann, wenn $TG < c_1$ bzw. $TG > c_2$ ist, also bei extremen Abweichungen der Testgröße in Richtung der Alternativhypothese. Die kritischen Werte c_1 bzw. c_2 sind ebenfalls im Anhang angeführt.

Bei großem Stichprobenumfang (als Faustregel gilt n ≥ 30) wird für den Vorzeichentest die standardnormalverteilte Testgröße

$$TG = \frac{D^+ - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = \frac{2D^+ - n}{\sqrt{n}}$$

verwendet, welche sich durch Approximation der Binomialverteilung durch die Normalverteilung ergibt. Die kritischen Werte sind in diesem Fall natürlich als entsprechende Quantile der Standardnormalverteilung zu bestimmen.

Beispiel (Fortsetzung): Kehren wir nochmals zum letzten Beispiel zurück. Die Vorgabe der Geschäftsleitung über die Wartezeit an den Kassen entspricht der Nullhypothese H_0: $\tilde{\mu} \le 3$. Wir führen also einen einseitigen Test auf Überschreitung durch. Der kritische Wert zum Stichprobenumfang n = 12 und zum Signifikanzniveau α = 0,05 beträgt laut Tabelle $c_2 = 9$. Da die Wartezeit jedoch in $D^+ = 10$ Fällen über 3 Minuten liegt, wird die Nullhypothese wegen $TG = D^+ = 10 > 9 = c_2$ verworfen. Die mittlere Wartezeit liegt also signifikant über dem angestrebten Wert von 3 Minuten.

Der Vorzeichentest wird häufig zum **Vergleich der Mediane mittels verbundener Stichproben**, also z.B. beim Vergleich zweier Produktionsverfahren oder Behandlungsmethoden, wo die Messwerte einander paarweise zugeordnet sind, angewendet und stellt damit eine nichtparametrische Alternative zum Differenzen-t-Test dar. Dabei bildet man die Differenzen einander entsprechender Werte und überprüft, ob diese aus einer Grundgesamtheit mit dem Median $\tilde{\mu} = 0$ (bzw. $\tilde{\mu} \geq 0$ oder $\tilde{\mu} \leq 0$) stammen. Testgröße ist dann die Anzahl D^+ der positiven Differenzen, die wieder binomialverteilt mit p = 1/2 ist. Daher rührt auch der Name Vorzeichentest.

Beispiel: Zur Erprobung eines neuen Düngemittels wurden 20 Anbaueinheiten je zur Hälfte nach einem herkömmlichen Verfahren bzw. zur Hälfte mit dem neuen Düngemittel behandelt, und die Erträge für beide Methoden miteinander verglichen. In 14 Fällen konnte eine Ertragssteigerung, in 5 Fällen ein Rückgang des Ertrags und in einem Fall keine Veränderung festgestellt werden. Spricht dieses Versuchsergebnis für das neue Düngemittel?

Zur Beantwortung dieser Frage wird angenommen, dass das neue Düngemittel keine Ertragsverbesserung mit sich bringt, sondern mit dem herkömmlichen Verfahren gleichwertig ist. In diesem Fall können für das Merkmal Ertragssteigerung im Durchschnitt gleich viele positive wie negative Werte erwartet werden. Scheidet man die eine neutrale Beobachtung aus, so verbleibt eine Stichprobe vom Umfang n = 19, darunter $D^+ = 14$ positive Werte. Die Tabelle für den zweiseitigen Vorzeichentest weist zum Signifikanzniveau $\alpha = 0{,}05$ einen Annahmebereich von 5 bis einschließlich 14 aus. Da die Testgröße $TG = D^+ = 14$ somit noch im Annahmebereich liegt, kann aus der vorliegenden Stichprobe nicht auf eine Ertragssteigerung durch das neue Düngemittel geschlossen werden.

Umfasst die Stichprobe jedoch 50 Anbaueinheiten, von denen in 35 Fällen eine Ertragssteigerung, in 13 Fällen ein Ertragsrückgang und in 2 Fällen keine Veränderung beobachtet wurden, also annähernd gleiche Proportionen wie oben vorliegen, dann sieht das Ergebnis schon anders aus: Wir scheiden wieder die neutralen Werte aus und berechnen aus der verbleibenden Stichprobe vom Umfang n = 48 mit $D^+ = 35$ positiven Werten die standardnormalverteilte Testgröße gemäß

$$TG = \frac{2D^+ - n}{\sqrt{n}} = \frac{2 \cdot 35 - 48}{\sqrt{48}} = 3{,}18 .$$

Wählen wir wieder $\alpha = 0{,}05$ und bestimmen den zugehörenden kritischen Wert c = 1,96 aus $\Phi(c) = 1 - \alpha/2 = 0{,}975$, so kann diesmal wegen TG > c die Nullhypothese gleicher Erträge klar verworfen werden, d.h., die Daten sprechen nun deutlich für eine Ertragssteigerung durch das neue Düngemittel.

Der Vorzeichentest ist zwar vollkommen unabhängig von der tatsächlichen Verteilung der Grundgesamtheit(en), er ist jedoch ein ziemlich grober Test. Gehen doch die einzelnen Stichprobenwerte überhaupt nicht in das Testverfahren ein, sondern lediglich die Information, wie viele Werte größer bzw. kleiner als eine vorgegebene Zahl sind. Ein Vergleich des Vorzeichentests mit dem t-Test führt, wie man zeigen kann, zu einer asymptotischen Effizienz von etwa 64%. Das bedeutet, dass man im Fall einer normalverteilten Grundgesamtheit beim t-Test mit einem Stichprobenumfang von 64 Werten dieselbe Güte erreicht wie beim Vorzeichentest mit 100 Werten.

(ii) Wilcoxon-Test als Beispiel für einen Rangtest

Im Gegensatz zum Vorzeichentest kann mehr Information aus der Stichprobe verwendet werden, wenn man auch die Reihenfolge der Stichprobenwerte berücksichtigt. Dieser Ansatz führt zu einer Gruppe von nichtparametrischen Testverfahren, den so genannten **Rangtests**.

Rangtests sind Verfahren, bei denen anstelle der einzelnen Stichprobenwerte nur deren Rangzahlen verwendet werden. Diese erhält man folgendermaßen. Gegeben sei z.B. eine Stichprobe mit n = 10 Beobachtungswerten:

$$x_1 = 21,\ x_2 = 26,\ x_3 = 26,\ x_4 = 16,\ x_5 = 23,\ x_6 = 42,\ x_7 = 29,\ x_8 = 26,\ x_9 = 23,\ x_{10} = 25.$$

Ordnet man sämtliche Stichprobenwerte x_i der Größe nach, so erhält man eine geordnete Stichprobe, deren Elemente wir mit $x_{(i)}$ bezeichnen:

$$x_{(1)} = 16,\ x_{(2)} = 21,\ x_{(3)} = 23,\ x_{(4)} = 23,\ x_{(5)} = 25,$$

$$x_{(6)} = 26,\ x_{(7)} = 26,\ x_{(8)} = 26,\ x_{(9)} = 29,\ x_{(10)} = 42.$$

Der Index i von $x_{(i)}$ gibt dabei den Platz an, den dieser Wert in der geordneten Stichprobe einnimmt, und wird als **Rangzahl** von $x_{(i)}$ bezeichnet. Insbesondere ist $x_{(1)}$ der kleinste und $x_{(n)}$ der größte Wert in der Stichprobe. Treten zwei oder mehrere gleich große Beobachtungswerte auf – man spricht in diesem Fall von **Bindungen** – dann ersetzt man sämtliche Rangzahlen, die für die gebundenen Werte insgesamt zu vergeben sind, durch ihr arithmetisches Mittel. Somit ergeben sich schließlich für die gegebene Stichprobe die folgenden Rangzahlen r_i:

Wert x_i	21	26	26	16	23	42	29	26	23	25
Rang r_i	2	7	7	1	3,5	10	9	7	3,5	5

Als Alternative zum Vorzeichentest zur Überprüfung der Lage einer Verteilung oder für den Lagevergleich zweier Verteilungen mittels verbundener Stichproben kann der **Wilcoxon-Test** als geeigneter Rangtest verwendet werden.

Der Wilcoxon-Test zur Überprüfung der Hypothese H_0: $\tilde{\mu} = \tilde{\mu}_0$ für den Median einer stetigen und symmetrischen Verteilung verläuft nach folgendem Schema: Den Ausgangspunkt bildet eine Zufallsstichprobe $x_1, ..., x_n$ aus der betrachteten Grundgesamtheit, wobei wiederum nur jene Werte berücksichtigt werden, die von $\tilde{\mu}_0$ verschieden sind. Dann bildet man die Differenzen $d_i = x_i - \tilde{\mu}_0$ (die sämtliche von 0 verschieden sind), ordnet sie nach ihrer absoluten Größe und ermittelt die Rangzahlen dieser absoluten Differenzen. Schließlich addiert man jene Ränge, welche zu positiven Differenzen d_i gehören, zur Rangsumme R^+ und berechnet damit die Teststatistik

$$TG = R^+ - \frac{n(n+1)}{4}.$$

Die Quantile der Teststatistik TG sind im Anhang tabelliert. Aus der Tabelle entnimmt man – beim einseitigen wie beim zweiseitigen Wilcoxon-Test – zu vorgegebenem Signifikanzniveau

α und Stichprobenumfang n den entsprechenden kritischen Wert c. Gilt dann $TG < -c$ oder $TG > c$, so wird die Nullhypothese H_0 verworfen, im Fall $-c \le TG \le c$ wird sie beibehalten.

Der Wilcoxon-Test berücksichtigt zwar mehr Informationen aus der Stichprobe als der gewöhnliche Vorzeichentest, er ist aber streng genommen nur unter der Voraussetzung einer symmetrischen Verteilung der Grundgesamtheit anwendbar. Seinem parametrischen Gegenstück, dem Einstichproben-t-Test, ist der Wilcoxon-Test stets dann vorzuziehen, wenn die Annahme einer normalverteilten Grundgesamtheit nicht gerechtfertigt erscheint.

Für Stichproben mit einem Umfang $n \ge 30$ kann die Verteilung der Rangsumme R^+ durch die Normalverteilung approximiert werden. In diesem Fall verwendet man die Teststatistik

$$TG = \frac{R^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}},$$

welche approximativ standardnormalverteilt ist.

Beispiel (Fortsetzung): Kommen wir nochmals auf das Beispiel zu den Wartezeiten an der Kassa eines Supermarkts zurück, und prüfen die Nullhypothese H_0: $\tilde{\mu} \le 3$ für die mittlere Wartezeit $\tilde{\mu}$ mit dem Wilcoxon-Test. Dazu bilden wir aus der Stichprobe der Wartezeiten x_i die Differenzen $d_i = x_i - 3$ für i = 1, ..., n = 12, ordnen diese nach ihrem Absolutbetrag und bestimmen die zugehörigen Rangzahlen. Diese Berechnungen sind in nachstehender Tabelle zusammengefasst:

Wartezeit x_i	5	4,5	3,5	6	1,5	12	15	3,5	5	7,5	2,5	4
Differenz d_i	2	1,5	0,5	3	-1,5	9	12	0,5	2	4,5	-0,5	1
Rang r_i	7,5	5,5	2	9	5,5	11	12	2	7,5	10	2	4

Die Summe der Rangzahlen zu den positiven Differenzen beträgt $R^+ = 70{,}5$ und die Teststatistik $TG = R^+ - n(n+1)/4 = 31{,}5$. Der Tabelle für den Wilcoxon-Test im Anhang entnehmen wir für den einseitigen Test mit $\alpha = 0{,}05$ und n = 12 den kritischen Wert c = 21,0. Wegen TG > c wird die Nullhypothese verworfen, d.h., die mittlere Wartezeit an der Kassa liegt entgegen den Vorgaben über 3 Minuten.

(iii) U-Test von Mann und Whitney

Als nichtparametrische Alternative zum klassischen Zweistichproben-t-Test stellen wir nun den **U-Test von Mann und Whitney** als bekanntesten Test für den Lagevergleich zweier beliebiger stetiger Verteilungen unter Benutzung unabhängiger Stichproben vor. Dazu gehen wir von zwei Verteilungen aus, die sich nicht in ihrer Form, sondern höchstens in ihrer Lage voneinander unterscheiden. Die Nullhypothese für den zweiseitigen Test lautet dann H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$ für die entsprechenden Mediane $\tilde{\mu}_1$ und $\tilde{\mu}_2$, die Alternative lautet H_1: $\tilde{\mu}_1 \ne \tilde{\mu}_2$. Die Entscheidung wird unter Benutzung zweier unabhängiger Stichproben von den Umfängen n_1 bzw. n_2 wie folgt getroffen:

Man ordnet alle $n_1 + n_2$ Stichprobenwerte der Größe nach, bildet die zugehörenden Rangzahlen und addiert sodann nur die Rangzahlen der ersten Stichprobe (in der kombinierten geordneten Stichprobe) zur Rangsumme R_1. Daraufhin wird die Teststatistik

$$TG = R_1 - \frac{n_1(n_1 + n_2 + 1)}{2}$$

berechnet. Die Herleitung der Verteilung von TG ist durch kombinatorische Überlegungen möglich, kritische Werte c für TG sind in Abhängigkeit vom Signifikanzniveau α für verschiedene Stichprobenumfänge n_1 und n_2 im Anhang tabelliert. Fällt TG in den kritischen Bereich, d.h. gilt $TG < -c$ oder $TG > c$, so wird die Nullhypothese verworfen, ansonsten wird sie beibehalten. Der U-Test kann natürlich auch als einseitiger Test geführt werden, entsprechende kritische Werte für den einseitigen Fall sind ebenfalls im Anhang angegeben.

Als Näherung für große Stichproben ($n_1, n_2 \geq 10$, also mindestens 10 pro Gruppe) wird die normalisierte Testgröße

$$TG = \frac{R_1 - n_1(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

verwendet, welche unter der Nullhypothese annähernd standardnormalverteilt ist.

Beispiel: Zum Vergleich von Kondensatoren aus zwei verschiedenen Lieferungen bezüglich ihrer Kapazität werden der ersten Lieferung eine Stichprobe vom Umfang $n_1 = 12$ und der zweiten Lieferung $n_2 = 9$ Kondensatoren zufällig entnommen. Die gemessenen Kapazitäten (in μF) sind in nachstehender Tabelle angegeben. Es soll untersucht werden, ob sich die beiden Lieferungen voneinander wesentlich unterscheiden, oder ob zwischen den Lieferungen kein Unterschied besteht.

	1. Lieferung ($n_1 = 12$)		**2. Lieferung ($n_2 = 9$)**	
Nummer	**Kapazität**	**Rang**	**Kapazität**	**Rang**
1	1,90	4	1,84	1
2	2,00	6	1,86	2
3	2,02	8	1,89	3
4	2,02	8	1,95	5
5	2,04	10	2,02	8
6	2,08	12	2,07	11
7	2,19	14	2,13	13
8	2,24	16	2,22	15
9	2,26	17	2,31	19
10	2,28	18		
11	2,35	20,5		
12	2,35	20,5		
		$R_1 = 154$		$R_2 = 77$

Zur Prüfung der Nullhypothese H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$ für die Mediane $\tilde{\mu}_1$ und $\tilde{\mu}_2$ der beiden Lieferungen ordnet man alle $n_1 + n_2 = 21$ Messwerte der Größe nach und vergibt die Rangzahlen von 1 bis 21 wie in der Tabelle angegeben. Die Summe der Rangzahlen aus der ersten Stichprobe beträgt $R_1 = 154$, und für die Testgröße errechnet man

$$TG = 154 - \frac{12 \cdot (12+9+1)}{2} = 22 .$$

In der Tabelle für den zweiseitigen U-Test findet man zur Signifikanzzahl $\alpha = 0{,}05$, $n_1 = 12$ und $n_2 = 9$ den kritischen Wert $c = 27$. Somit verläuft der Annahmebereich von -27 bis 27, und wegen $-c \le TG \le c$ wird die Nullhypothese beibehalten, d.h., die Kapazitäten der Kondensatoren aus der ersten Lieferung sind nicht wesentlich größer als jene der zweiten Lieferung.

Da obige Fragestellung natürlich symmetrisch ist, hätten wir eine Testgröße genauso gut mittels der Rangsumme R_2 aus den Rangzahlen der zweiten Lieferung bilden können. Dabei ergibt sich

$$TG_2 = R_2 - \frac{n_2(n_1+n_2+1)}{2} = 77 - \frac{9 \cdot (12+9+1)}{2} = -22$$

und folglich dieselbe Entscheidung für die Beibehaltung der Nullhypothese, dass die beiden Lieferungen übereinstimmen. Ganz allgemein besteht zwischen den Rangsummen R_1 und R_2 der Zusammenhang

$$R_1 + R_2 = \sum_{k=1}^{n_1+n_2} k = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$$

und damit folgt $TG_1 + TG_2 = 0$, also $TG_2 = -TG_1$ für die beiden entsprechenden Teststatistiken TG_1 und TG_2.

Eine kurze Rechnung zeigt, dass dasselbe Ergebnis auch mit dem t-Test erzielt wird. Bemerkenswert ist die hohe Effizienz des U-Tests im Vergleich mit seinem parametrischen Gegenstück, dem Zweistichproben-t-Test. Bei normalverteilten Grundgesamtheiten und einem Stichprobenumfang von insgesamt 100 Beobachtungen ist der U-Test etwa genauso gut, d.h., er besitzt dieselbe Gütefunktion, wie ein t-Test mit 95 Beobachtungswerten, er besitzt also eine asymptotische Effizienz von 95%. Bedenkt man ferner, dass der U-Test unter viel schwächeren Voraussetzungen durchführbar ist als der t-Test und zudem rechnerisch verhältnismäßig einfach ist, so lässt sich daran der große Vorteil nichtparametrischer Methoden ermessen.

Als Verallgemeinerungen des U-Tests für den Vergleich von mehr als zwei Verteilungen können der **H-Test von Kruskal und Wallis** sowie der **Friedmann-Test** verwendet werden. Beide Tests zählen zur Gruppe der Rangtests und können als nichtparametrisches Gegenstück zur Varianzanalyse angesehen werden.

Die Chi-Quadrat-Methode

Die bisher behandelten Tests ermöglichen einen Vergleich von Lageparametern mit vorgegebenen Werten oder auch untereinander. In diesem Abschnitt wird nun ein Verfahren beschrieben, mit dessen Hilfe eine Verteilung in ihrem Gesamtverlauf beurteilt werden kann.

Der älteste und wohl auch bekannteste Test zum Vergleich von empirischen und theoretischen Verteilungen ist der von Pearson eingeführte **χ^2-Test**. Die Grundlage dieses Tests bildet die χ^2-Verteilung, die wir bereits kennen gelernt haben.

(iv) χ^2-Anpassungstest: Vergleich von beobachteten und erwarteten Häufigkeiten

Wir betrachten eine diskrete Verteilung, bei der genau k Ausprägungen a_1, a_2, ..., a_k mit positiver Wahrscheinlichkeit auftreten können. Aus dieser Verteilung sei eine Stichprobe vom Umfang n gegeben, wobei jeweils n_i Stichprobenelemente von der Ausprägung a_i sind (i = 1, 2, ..., k). Die Stichprobe ist somit in k Klassen eingeteilt. Die absoluten Klassenhäufigkeiten betragen n_1, n_2, ..., n_k und es gilt $n_1 + n_2 + ... + n_k = n$.

Getestet wird nun eine Hypothese über die zugrunde liegende Wahrscheinlichkeitsverteilung, nämlich H_0: Die Wahrscheinlichkeiten der Ausprägungen a_1, a_2, ..., a_k sind p_1, p_2, ..., p_k (mit $p_1 + p_2 + ... + p_k = l$). Um zu einem Urteil über die Richtigkeit dieser Hypothese zu gelangen, müssen die tatsächlich **beobachteten** Werte n_1,..., n_k mit den entsprechenden unter H_0 **erwarteten** Häufigkeiten $n_1^* = np_1$, ..., $n_k^* = np_k$ verglichen werden. Dieser Vergleich verläuft nach folgendem Schema:

Zum Test der Hypothese H_0, die gegebene Stichprobe entstamme einer durch die Wahrscheinlichkeiten p_i charakterisierten Verteilung, bildet man die Testgröße

$$TG = \sum_{i=1}^{k} \frac{(n_i - n_i^*)^2}{n_i^*},$$

wobei n_i die in der Stichprobe beobachteten, $n_i^* = np_i$ die unter der Nullhypothese erwarteten Häufigkeiten darstellen. Die Verteilung der Testgröße TG kann für große Stichproben annähernd durch eine χ^2-Verteilung mit k − 1 Freiheitsgraden beschrieben werden. Der Test wird als einseitiger Test auf Überschreitung geführt, so dass große Werte für TG signifikant sind. Die Testentscheidung wird daher nach dem folgenden Verfahren getroffen: Man wählt eine Signifikanzzahl α und bestimmt einen entsprechenden kritischen Wert c für die χ^2-Verteilung mit FG = k − 1 gemäß $F(c) = 1 - \alpha$. Der kritische Bereich ist dann durch TG > c festgelegt. Gilt also TG > c, so wird die Nullhypothese verworfen, ist TG ≤ c, wird sie beibehalten.

Beispiel: Mit einem Würfel wird 60 mal gewürfelt. Dabei werden die folgenden Häufigkeiten für die Augenzahlen 1 bis 6 beobachtet: $n_1 = 7$, $n_2 = 8$, $n_3 = 11$, $n_4 = 5$, $n_5 = 10$ und $n_6 = 19$. Ist der Würfel fair, d.h. gilt für die entsprechenden Wahrscheinlichkeiten $p_1 = ... = p_6 = 1/6$?

Wir haben es hier mit einer Stichprobe vom Umfang n = 60 aus einer diskreten Grundgesamtheit mit k = 6 Ausprägungen zu tun. Die Nullhypothese lautet H_0: $p_1 = ... = p_6 = 1/6$. Wir wählen das Signifikanzniveau $\alpha = 0{,}05$ und bestimmen den kritischen Wert c aus der Gleichung $F(c) = 1 - \alpha = 0{,}95$ für die χ^2-Verteilung mit FG = k − 1 = 5 Freiheitsgraden. Dabei ergibt sich c = 11,07 (laut Tabelle im Anhang).

Bei Gültigkeit der Hypothese H_0, d.h. falls der Würfel fair ist, sind unter den n = 60 Würfen jeweils die Häufigkeiten $n_i^* = np_i = 10$ für alle i zu erwarten. Damit berechnen wir den Wert der Testgröße

$$TG = \frac{(7-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(19-10)^2}{10} = 12 .$$

Aus TG > c ersehen wir, dass H_0 auf dem 5%-Niveau verworfen werden muss, d.h., dass der Würfel nicht mehr als fair angesehen werden kann.

Wie bereits erwähnt, stellt der χ^2-Test lediglich ein Näherungsverfahren dar. Wie gut die Verteilung von TG mit der χ^2-Verteilung übereinstimmt, hängt vor allem von den erwarteten Häufigkeiten n_i^* in den schwach besetzen Klassen ab. Als Faustregel gilt $n_i^* \geq 5$ für mindestens 80% aller Klassen und zugleich $n_i^* \geq 1$ für alle Klassen. Ist diese Voraussetzung nicht erfüllt, müssen einzelne Klassen zusammengelegt und die Anzahl der Freiheitsgrade entsprechend reduziert werden.

Obwohl die Fragestellung bei diesem Testverfahren eine zweiseitige Problemstellung darstellt, ist der Annahmebereich einseitig: Die Nullhypothese wird nur dann verworfen, wenn die Teststatistik TG den kritischen Wert c übersteigt. Der Grund dafür liegt darin, dass Abweichungen zwischen den beobachteten und den erwarteten Häufigkeiten in jeder Richtung stets zu einer Vergrößerung von TG führen. Abschließend sei auch darauf verwiesen, dass zur Berechnung der Prüfgröße die absoluten (und nicht die relativen) Häufigkeiten herangezogen werden müssen; ein Vergleich von relativen Häufigkeiten mittels der Formel für TG ist nicht zulässig.

Der χ^2-Test tritt in der mathematischen Statistik in unterschiedlichen Zusammenhängen auf. Allgemein kann der Test in der besprochenen Form als Test für die Güte der Anpassung (**goodness of fit**) an eine gegebene Verteilungsfunktion verwendet werden. Bezeichnet man die Verteilungsfunktion der Grundgesamtheit, aus welcher die Stichprobe stammt, mit F und die Verteilungsfunktion der vorgegebenen Verteilung mit F_0, so kann man die Nullhypothese des Anpassungstests kurz in der Form H_0: $F = F_0$ schreiben. Sind die beiden betrachteten Verteilungen diskret, so ist eine Gruppierung der Daten bereits vorgegeben, und der Anpassungstest kann durch Vergleich der aus der Verteilung F beobachteten und der gemäß F_0 erwarteten Häufigkeiten wie beschrieben durchgeführt werden. Bei stetigen Verteilungen hingegen muss durch eine geeignete Klassenbildung erst eine Gruppierung geschaffen werden. Mit dem χ^2-Test können dann wieder die in der Stichprobe beobachteten Klassenhäufigkeiten mit den absoluten Häufigkeiten jener Werte, die unter Annahme der Nullhypothese in jeder Klasse zu erwarten sind, verglichen werden. Ein weiteres Verfahren zur Prüfung der Übereinstimmung von Verteilungen stellt der **Kolmogoroff-Smirnoff-Test** dar.