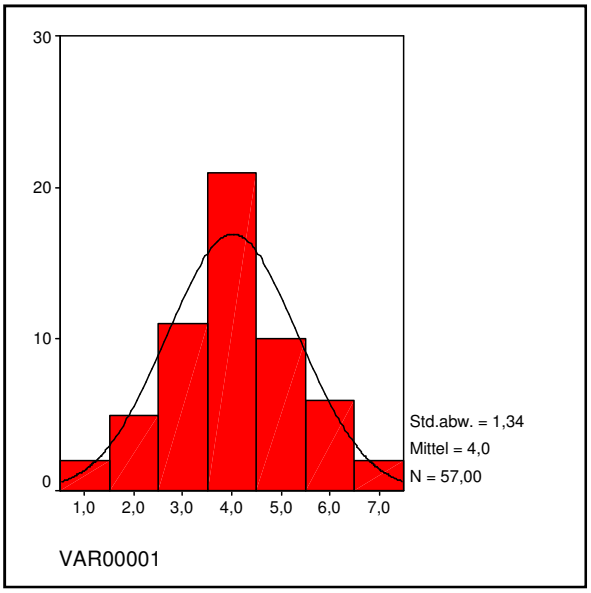
2. INFERENZSTATISTIK

„Inferenzstatistik“ bedeutet übersetzt „schließende Statistik“. Damit ist der Schluss von den erhobenen Daten einer Stichprobe auf Werte in der Population gemeint.

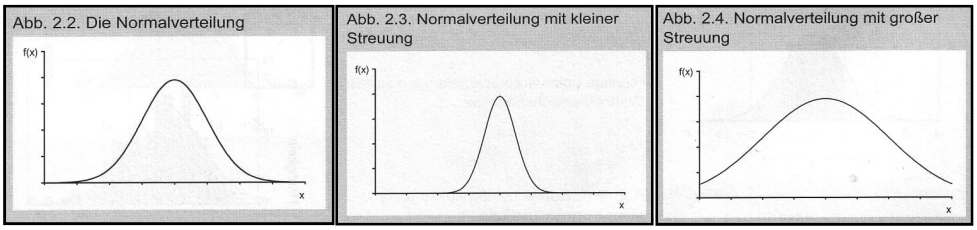
2.1 Die Normalverteilung

In der Natur sind sehr viele Merkmale normalverteilt. Dies gilt beispielsweise für die Körpergröße, Intelligenz oder das Sehvermögen. Die Normalverteilung galt sogar zeitweise als eine Art Naturgesetz. Auch wenn diese Vorstellung mittlerweile abgelehnt wird, so zeigt sie doch die herausragende Bedeutung der Normalverteilung in der Mathematik und in den empirischen Sozialwissenschaften. Sehr viele statistische Verfahren, die für wissenschaftliches Arbeiten erforderlich sind, setzen voraus, dass die relevanten Merkmale normalverteilt sind.



Bei der Abbildung 2.1 handelt es sich um eine diskrete Wahrscheinlichkeitsverteilung. Sie zeichnet sich dadurch aus, dass:

- alle Werte auf der x-Achse getrennt voneinander stehen
- jedem einzelnen Wert eine bestimmte Wahrscheinlichkeit zugeordnet ist



Kontinuierliche Wahrscheinlichkeitsverteilungen zeichnen sich dadurch aus, dass:

- der Abstand der Werte auf der x-Achse unendlich klein ist
- einem einzelnen Werte keine bestimmte Wahrscheinlichkeit zugeordnet werden kann

Ein anderes Beispiel für eine kontinuierliche Variable ist das Gewicht. Zwar wird es zumeist in ganzen Zahlen angegeben, tatsächlich nimmt es aber unendlich viele verschiedene Ausprägungen zwischen diesen ganzen kg-Werten an, z.B. 8,657 kg.

Das Resultat dieses Gedankenexperimentes ist eine unimodale und eingipflige Verteilung mit glockenförmigen Verlauf. Sie ist symmetrisch und nähert sich der x-Achse asymptotisch an. Dadurch sind die Werte für Median, Modus und arithmetisches Mittel identisch. Eine solche Verteilung heißt Normalverteilung.

Der Entdecker der Normalverteilung heißt Carl Friedrich Gauss. Sein Portrait war auf jedem Zehn-Mark-Schein abgebildet. Dort fand sich auch die mathematische Funktion der Normalverteilung:

$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Jede Normalverteilung ist durch ihr arithmetisches Mittel und ihre Streuung vollständig determiniert.

2.1.2 Wahrscheinlichkeiten unter der Normalverteilung

Aus der Symmetrie der Normalverteilung folgt, dass die Wahrscheinlichkeit, einen größeren Wert als den Mittelwert zu messen, p: 0,5 ist.

Da eine Kurve aus unendlich vielen Punkten besteht, ist die Wahrscheinlichkeit für einen einzelnen Wert unendlich klein. Deshalb sind Wahrscheinlichkeitsberechnungen nur für Flächen, also Intervalle unter der Kurve möglich, niemals für einzelne Werte.

Für Normalverteilungen gilt, dass die Fläche, die von +/- einer Standardabweichung vom Mittelwert begrenzt wird, mehr als 2/3 aller (68,26%) beinhaltet. 95,44% liegen im Bereich von +/- zwei Standardabweichungen.

2.2 Die Standardnormalverteilung

Unter den unendlich vielen Normalverteilungen gibt es eine mit dem Mittelwert $\mu = 0$ und der Streuung $\sigma = 1$. Diese wird als Standardnormalverteilung bezeichnet. Ihr kommt eine besondere Bedeutung zu, denn jede erdenkliche Normalverteilung ist durch eine einfache Transformation in diese standardisierte Form zu überführen. Dies ist die in Kapitel 1.4 behandelte z-Transformation nach der Formel:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Standardnormalverteilung und z-Verteilung sind also Bezeichnungen für ein- und dieselbe Verteilung. Deren Hochpunkt befindet sich bei 0, die Wendepunkte bei –1 und +1. Da die Streuung der Standardnormalverteilung σ = 1 beträgt, lassen sich z-Werte hier direkt als Standardabweichungseinheiten vom Mittelwert auffassen.

Die Standardnormalverteilung macht den Sozialwissenschaftlern das Leben in vielen Bereichen leichter. Mit ihrer Hilfe und den eng mit ihr verknüpften z-Werten ist es problemlos möglich, die Ergebnisse mehrerer, auf unterschiedlichen Normalverteilungen basierender Messinstrumente zu vergleichen. Dies ist z.B. dann nötig, wenn untersucht werden soll, ob zwei verschiedene psychologische Tests wirklich dasselbe Konstrukt messen oder wenn Umfragedaten fremder Institute, die mit unterschiedlichen Instrumenten arbeiten, untereinander verglichen werden sollen.

2.3 Die Stichprobenkennwerteverteilung

Nehmen wir einmal an, statt der einen hätten wir theoretisch unendlich viele, gleich große, voneinander unabhängige Stichproben erheben können, die wiederum unendlich viele, unterschiedlich große Stichprobenmittelwerte liefern.

Die resultierende Verteilung dieser Mittelwerte ginge mit steigender Anzahl von Werten in eine Normalverteilung über, die sogenannte Stichprobenkennwerteverteilung von Mittelwerten. Sie umfasst alle möglichen Mittelwerte, die aus einer Stichprobe der gegebenen Größe entstehen können. Der Mittelwert dieser Stichprobenkennwerteverteilung von Mittelwerten $(\bar{\bar{x}})$ repräsentiert die Verteilung am besten, denn alle Mittelwerte streuen um ihn

herum (Kap.2.1). Deshalb ist bei einer einmaligen Ziehung einer Stichprobe ein Mittelwert in der Nähe von $\bar{\bar{x}}$ wahrscheinlicher als ein Mittelwert, der sehr weit davon entfernt ist. In der Sprache der Statistik heißt $\bar{\bar{x}}$ deshalb auch Erwartungswert von $\bar{x}(E(\bar{x}))$. Man kann zeigen, dass dieser Erwartungswert gleich dem Populationsmittelwert μ ist. Es gilt also $E(\bar{x}) = \mu$.

Der Standardfehler des Mittelwerts

Die Streuung der Stichprobenkennwerteverteilung heißt Standardfehler des Mittelwerts. Mit seiner Hilfe lässt sich die Genauigkeit der Schätzung des Populationsmittelwertes beurteilen. Er ist definiert als die Streuung in einer Verteilung von Mittelwerten aus gleich großen Zufallsstichproben einer Population.

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma_x^2}{n}} = \frac{\sigma_x}{\sqrt{n}}$$

Der Standardfehler wird zwangsläufig umso größer, je größer auch die Streuung der Messwerte in der Population ist. Die Populationsstreuung steht in der Formel im Zähler.

Beispiel: Theoretisch könnten alle Menschen einen Intelligenzquotienten von exakt 100 haben. Da die Intelligenz in der Bevölkerung aber eine gewisse Varianz (und damit Streuung) aufweist, wird ihr Populationsmittelwert zwar erwartungstreu geschätzt, den exakten Wert wird aber nie jemand wissen.

Wenn die Größe des Standardfehlers entscheidend die Güte der Mittelwertsschätzung beeinflusst, ist es von Vorteil, auch den Standardfehler zu schätzen. Wie oben ausgeführt ist er proportional zur Populationsstreuung und verringert sich bei zunehmendem Stichprobenumfang. Mathematisch berechnet er sich wie folgt:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}} = \sqrt{\frac{\hat{\sigma}_x^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n \cdot (n-1)}}$$

Noch einmal: Ein Mittelwert stellt eine umso präzisere Schätzung des Populationsparameters μ dar, je kleiner sein Standardfehler ist.

Ein großer Vorteil ist, dass die Daten einer Stichprobe ausreichen, um die Größe des Standardfehlers zu schätzen. Auf Grund einer Stichprobe können wir deshalb nicht nur den Mittelwert der Population schätzen, sondern mit Hilfe des Standardfehlers auch die Genauigkeit dieser Schätzung bestimmen.

Mit einer Wahrscheinlichkeit von 68,26% liegt der wahre Populationsmittelwert zwischen +/- einem Standardfehler. Ein solcher Bereich heißt Vertrauensintervall oder Konfidenzintervall des Mittelwerts. Es ist ein Maß für die Präzision, mit der ein Stichprobenkennwert den „wahren“ Populationsparameter schätzt.

In der Praxis wird per Konvention meist ein 95% oder 99% Konfidenzintervall bestimmt.
Da ein Konfidenzintervall symmetrisch um einen Mittelwert konstruiert wird, müssen wir auf den beiden Seiten der Verteilung jeweils 2,5% bzw. 0,5% abschneiden, um einen Wahrscheinlichkeitsbereich von 95% oder 99% zu erhalten. Ein z-Wert von –1,96 (-2,58) schneidet 2,5% (0,5%) nach links ab, ein Wert von 1,96 (2,58) schneidet denselben Bereich nach rechts ab. Aufgrund der Symmetrie der z-Verteilung genügt es, den Betrag von einem der beiden z-Werte zu bestimmen. Dieser Betrag muss mit dem Standardfehler des Mittelwerts multipliziert werden. Das Ergebnis wird zum Mittelwert addiert bzw. vom Mittelwert subtrahiert, um die beiden Grenzen des Konfidenzintervalls zu erhalten:

untere Grenze = $\bar{x} - z \cdot \sigma_{\bar{x}}$

obere Grenze = $\bar{x} + z \cdot \sigma_{\bar{x}}$

Für ein 95% Konfidenzintervall: z: 1,96

Für ein 99% Konfidenzintervall: z: 2,58

Bei kleinen Stichproben (N<30) ist allerdings die Normalverteilungsannahme häufig verletzt (siehe Bortz, 2005, S. 103). Das korrekte Konfidenzintervall ist in diesem Fall etwas größer als die Berechnung über z-Werte angibt. Hier empfiehlt es sich, anstatt des z-Wertes den zugehörigen Wert der t-Verteilung (bei df = n-1) zu ermitteln, der 2,5% bzw. 0,5% nach oben bzw. unten abschneidet.

3. Der t-Test

Der t-Test untersucht, ob sich zwei empirisch gefundene Mittelwerte systematisch voneinander unterscheiden. Mit Hilfe dieses Testverfahrens ist es möglich festzustellen, ob zwei betrachtete Gruppen in einem untersuchten Merkmal wirklich einen Unterschied aufweisen.

3.1 Was ist der t-Test?

3.1.1 Die Fragestellung des t-Tests

Der t-Test liefert nur für intervallskalierte Daten zuverlässige Informationen. Deshalb gehört er zur Gruppe der parametrischen Verfahren.
Parametrische Verfahren schätzen Populationsparameter mittels statistischer Kennwerte wie dem arithmetischen Mittel oder der Varianz, für deren Berechnung die Intervallskaliertheit der Daten Voraussetzung ist.

Er liefert eine Entscheidungshilfe dafür, ob ein gefundener Mittelwertsunterschied rein zufällig entstanden ist, oder ob es wirklich bedeutsame Unterschiede zwischen den zwei untersuchten Gruppen gibt.

Der wichtigste Wert für die Durchführung eines t-Tests ist die Differenz der Gruppenmittelwerte. Diese Differenz bildet den Stichprobenkennwert des t-Tests:

$$\bar{x}_1 - \bar{x}_2$$

Die zentrale Frage des t-Tests lautet: Wie wahrscheinlich ist die empirisch gefundene oder eine größere Mittelwertsdifferenz unter allen möglichen rein theoretisch denkbaren Differenzen?

Der t-Test dient wie viele andere statistische Verfahren zur Überprüfung aufgestellter Hypothesen. Dabei ist es wichtig, vor der Durchführung eines t-Tests die zu untersuchende Hypothese inhaltlich zu präzisieren.

Der t-Test kann jeweils nur zwei Gruppen im Mittelwertsvergleich betrachten.
Inhaltliche Hypothesen müssen in statistische Hypothesen umformuliert werden.

Bsp. $\bar{x}_1 - \bar{x}_2 > 0$. Die Bildung der Differenz wird entscheidend durch die Formulierung der statistischen Hypothese mitbestimmt: Sie legt fest, welcher Wert vom anderen abgezogen wird.

3.1.2 Die Nullhypothese

Für die Erklärung der Mittelwertsdifferenz gibt es neben der Annahme eines systematischen Unterschieds zwischen den beiden Gruppen eine weitere Möglichkeit: Die Differenz zwischen den Mittelwerten ist zufällig zustande gekommen und es gibt keinen echten Unterschied zwischen den beiden untersuchten Gruppen. Die beiden Gruppen stammen im Grunde aus zwei Populationen mit demselben Mittelwert. Die Differenz zwischen den Gruppen sollte demzufolge Null betragen. Diese Annahme heißt deshalb Nullhypothese oder H_0.

Stichprobenkennwerteverteilung unter der Nullhypothese

Unter Annahme der Nullhypothese kann eine Stichprobenkennwerteverteilung von Mittelwertsdifferenzen konstruiert werden. Alle möglichen zwei Stichprobenmittelwerte, aus denen die Differenzen gebildet werden, stammen unter Annahme der Nullhypothese aus zwei Populationen mit identischem Populationsmittelwert.

Wird aus zwei Populationen mit identischem Populationsmittelwert jeweils eine Stichprobe gezogen, so kann die Differenz der beiden Stichprobenmittelwerte theoretisch jeden beliebigen Wert annehmen. Die zu erwartende Differenz aber ist gleich Null, denn die Stichprobenmittelwerte sind normalverteilt um ihren jeweiligen Erwartungswert, den Populationsmittelwert.

Für die Bestimmung der Stichprobenkennwerteverteilung ohne Simulation muss ihre Streuung (Standardfehler der Mittelwertsdifferenz) mit Hilfe der Stichprobe geschätzt werden. In die Formel gehen die Stichprobenumfänge der betrachteten Gruppen und die geschätzten Streuungen der zugehörigen Populationen ein.

Die Formel lautet:

$$\hat{\sigma}_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} - \frac{\hat{\sigma}_2^2}{n_2}}$$

$\hat{\sigma}_{\bar{x}_1-\bar{x}_2}$: geschätzter Standardfehler der Mittelwertsdifferenz

n : Anzahl der Vpn bzw. Beobachtungen in der jeweiligen Stichprobe

$\hat{\sigma}$: geschätzte Varianz der jeweiligen Population

3.1.3 Die t-Verteilung

Die Standardisierung der Stichprobenkennwerteverteilung erfolgt ähnlich wie bei den z-Werten an ihrer geschätzten Streuung. Die standardisierten Stichprobenkennwerte heißen t-Werte, die standardisierten Verteilungen sind die t-Verteilungen (im Englischen auch „Student t“ genannt). Sie entsprechen nicht ganz der Standardnormalverteilung, sondern sind schmalgipfliger.

In einer t-Verteilung sind die Wahrscheinlichkeiten für die einzelnen t-Werte genau ablesbar. Die allgemeine Definition des t-Wertes lautet:

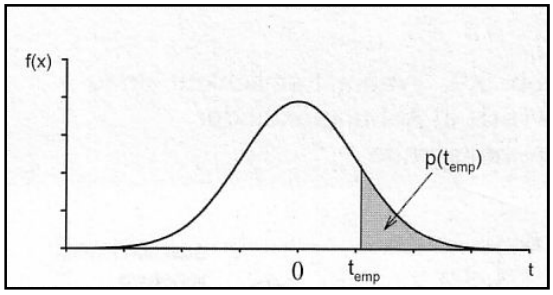
$$t_{df} = \frac{empirische\_Mittelwertsdifferenz - theoretische\_Mittelwertsdifferenz}{geschätzter\_S\,\text{tan}\,dardfehler\_der\_Mittelwertsdifferenz}$$

$$\text{formal: } t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

Der t-Test findet in den meisten Fällen als Nullhypothesensignifikanztest Anwendung. Diesem t-Test liegt die Annahme zu Grunde, dass die Populationsmittelwerte der beiden zu vergleichenden Gruppen identisch sind. Die theoretische Mittelwertsdifferenz unter der Nullhypothese kann bei der Berechnung weggelassen werden. Die vereinfachte Formel lautet:

$$t_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

Die Auftretenswahrscheinlichkeit eines positiven t-Wertes entspricht dem Anteil der Fläche unter der Kurve, den der t-Wert nach rechts abschneidet.



3.1.4 Die Freiheitsgrade einer t-Verteilung

Die exakte Form der t-Verteilung ist trotz der Standardisierung weiterhin vom Stichprobenumfang abhängig und deckt sich aus diesem Grunde nicht exakt mit der z-Verteilung. Der Unterschied zwischen diesen Verteilungen ist dadurch zu erklären, dass in die Berechnung des t-Werts nicht einer, sondern zwei erwartungstreue Schätzer für Populationsparameter eingehen: die empirische Mittelwertsdifferenz und der geschätzte Standardfehler der Mittelwertsdifferenz (in der Formel zur Berechnung der z-Werte ist die

Streuung kein Schätzer der Populationsstreuung, sondern bezieht sich direkt auf die Population).

Die Freiheitsgrade der gefundenen Mittelwertsdifferenz erlauben eine genaue Beschreibung der zu verwendenden t-Verteilung. Sie werden in dem hier besprochenen t-Test durch folgende Formel berechnet:

$$df = n_1 + n_2 - 2$$

3.1.5 Exkurs: Das Konzept der Freiheitsgrade

Die Genauigkeit, mit der Stichprobenkennwerte Populationsparameter schätzen, ist von der Anzahl ihrer Freiheitsgrade abhängig. Dadurch beeinflussen die Freiheitsgrade auch die Form solcher Verteilungen, in deren Berechnung geschätzte Größen eingehen.

Die Anzahl der Freiheitsgrade gibt an wie viele Werte in einer Berechungsformel frei variieren dürfen, damit es zu genau einem bestimmten Ergebnis kommt. So erlaubt eine Summe mit n Summanden die freie Wahl von n-1 Summanden, d.h. nur ein Summand ist aufgrund des Ergebnisses festgelegt.

Beispiel: In der Gleichung $x_1 + x_2 + x_3 = 15$ können für x_1 und x_2 beliebige Zahlen eingesetzt werden (z.B. 10 und 2), x_3 ist damit allerdings bereits bestimmt:

$10 + 2 + x_3 = 15 \qquad x_3 = 15 - 12 = 3.$

3.1.6 Bewertung des t-Werts

Die Nullhypothese nimmt an, dass der gefundene Unterschied der Mittelwerte zufällig zustande gekommen ist und die Stichproben aus zwei Populationen mit identischem Mittelwert stammen. Unter dieser Annahme errechnet der t-Test eine Wahrscheinlichkeit für das Auftreten der gefundenen oder einer größeren Differenz, die z.B. $p = 0{,}03$ beträgt.

Der errechnete Wert von 3% bedeutet, dass die Wahrscheinlichkeit für das Finden einer solchen oder einer größeren Differenz beim Ziehen von Stichproben aus einer identischen Population sehr gering ist. Natürlich ist diese Differenz möglich, sie ist aber sehr unwahrscheinlich. Wenn die Differenz unter Annahme der Nullhypothese sehr unwahrscheinlich ist, so trifft möglicherweise die Annahme selbst gar nicht zu.

Wenn die Annahme der Nullhypothese falsch ist und der Unterschied nicht auf Zufall beruht, dann muss die gefundene Mittelwertsdifferenz auf einem systematischen Unterschied zwischen den beiden Gruppen beruhen. Die Stichproben stammen dann nicht aus Populationen mit identischen, sondern mit verschiedenen Mittelwerten.

3.1.7 Entwicklung eines Entscheidungskriteriums

Wird die Nullhypothese abgelehnt, obwohl sie in Wirklichkeit gilt (also in der Population zutrifft), so gilt dies als α-Fehler.

Welche Wahrscheinlichkeit einer Fehlentscheidung ist bei Ablehnung der Nullhypothese tolerierbar? Es ist die Festlegung einer Grenze für die Ablehnung erforderlich. Eine solche Entscheidungsgrenze heißt Signifikanzniveau oder auch α-Fehler-Niveau. Ist die errechnete Auftretenswahrscheinlichkeit der Mittelwertsdifferenz kleiner als das Signifikanzniveau, so erfolgt die Ablehnung der Nullhypothese. Das Ergebnis wird als signifikant bezeichnet.

Die Wahl der Entscheidungsgrenze ist rein willkürlich und wird nur durch inhaltliche Überlegungen beeinflusst. Je nach Fragestellung kann es sinnvoll sein, ein relativ hohes (liberales) oder aber ein sehr strenges (konservatives) Signifikanzniveau zu wählen. Um diese Entscheidung kompetent treffen zu können, ist eine genaue Kenntnis der Materie und des betreffenden Forschungsstandes notwendig. Das Signifikanzniveau muss vor der Durchführung einer Untersuchung festgelegt und begründet werden, um das Ergebnis beurteilen zu können. Per Konvention liegt es meist bei $\alpha = 0{,}05$, einige Untersuchungen verwenden 1% oder 10%. Ein auf dem 5%-Niveau signifikantes Ergebnis wird in der Literatur und in SPSS in der Regel mit einem Sternchen (*) gekennzeichnet, ein auf dem 1%-Niveau signifikantes Ergebnis mit zwei Sternchen (**).

3.1.8 Population und Stichprobe beim t-Test

Der t-Test versucht, anhand einer empirischen Mittelwertsdifferenz zweier Stichproben auf die Größe der Differenz zwischen zwei Populationsmittelwerten zu schließen.

Um die Wahrscheinlichkeit der empirischen Mittelwertsdifferenz unter der Nullhypothese bestimmen zu können, ist eine Standardisierung der Stichprobenkennwerteverteilung an ihrer eigenen Streuung (dem geschätzten Standardfehler der Mittelwertsdifferenz) erforderlich, da es je nach Größe der Streuung unendlich viele Stichprobenkennwerteverteilungen gibt. Diese Verteilungen unterscheiden sich stark in ihrer Form. Die Standardisierung führt sie alle auf eine bestimmte Verteilung zurück, die t-Verteilung. Die Wahrscheinlichkeiten in einer t-Verteilung sind bekannt, hängen nur von den Freiheitsgraden ab und sind in Tabellen verzeichnet (siehe Anhang).

An dieser Stelle folgt der Schluss von der Stichprobe auf die Population: Ist der empirische t-Wert unter der Annahme der Nullhypothese sehr unwahrscheinlich, so ist diese theoretische

Annahme über die Populationsmittelwerte wahrscheinlich falsch. Also sind die Populationsmittelwerte wahrscheinlich nicht gleich, sondern verschieden.

3.1.9 Voraussetzungen für die Anwendung eines t-Tests

Für den t-Test gibt es drei mathematische Voraussetzungen:

1.) Das untersuchte Merkmal ist intervallskaliert
2.) Das untersuchte Merkmal ist in der Population normalverteilt.
3.) Die Populationsvarianzen, aus denen die beiden Stichproben stammen, sind gleich (Varianzhomogenität)

Um dies sicherzustellen ist es wichtig, dass die Stichproben der beiden Gruppen annähernd dieselbe Größe haben und nicht zu klein sind ($n_1 = n_2 > 30$). Erst wenn die Stichproben kleiner oder deutlich unterschiedlich groß sind, ist das Ergebnis eines t-Tests bei Verletzung der Voraussetzungen fehlerhaft.

Für den hier dargestellten t-Test für unabhängige Stichproben ist zusätzlich die Unabhängigkeit der Gruppen notwendig.

Testen der Voraussetzungen des t-Tests

Ein Test auf Normalverteilung (z.B. Kolmogorov-Smirnov-Test) ist bei kleinen Stichproben aufgrund der geringen Power nicht zu empfehlen. Auf eine Verletzung der Normalverteilungsannahme reagiert der t-Test ohnehin äußerst robust, eine ungefähre Symmetrie der Verteilung des Merkmals in der Population reicht aus, um eine annähernd normalverteilte Stichprobenkennwerteverteilung zu erzeugen.

In den meisten Fällen genügt eine grobe, deskriptive Kontrolle auf Normalverteilung. Aus diesen Gründen findet nur der Levene-Test der Varianzgleichheit häufiger Anwendung. Er vergleicht die Größe der Varianzen der zwei Gruppen: Der Test wird signifikant, wenn eine Varianz überzufällig größer ist als die andere.

3.2 Die Alternativhypothese

Die Alternativhypothese geht davon aus, dass die Populationen, aus denen die Stichproben gezogen werden, einen unterschiedlichen Populationsmittelwert haben. In ihrer allgemeinen Form umfasst sie alle möglichen Hypothesen, die nicht der Nullhypothese entsprechen. Diese Annahme wird formal wie folgt ausgedrückt:

$H_1 : \neg H_0$ (d.h. H_1 ist alles das, was H_0 nicht ist.)

3.2.1 Ungerichtete Hypothesen

Eine ungerichtete Alternativhypothese nimmt lediglich an, dass die Differenz der Populationsmittelwerte nicht gleich Null ist. Die Differenz kann also sowohl kleiner als auch größer Null sein. Man spricht von einer zweiseitigen Fragestellung. Die Nullhypothese beschränkt sich hier nur auf den Fall, in dem die Differenz Null ist. Die korrekte Schreibweise dieses Hypothesenpaares lautet:

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$

Das Signifikanzniveau liegt bei ungerichteten Fragestellungen ebenfalls meistens bei 5%. Allerdings verändert die Verwendung von ungerichteten Fragestellungen die Bestimmung der Signifikanz eines empirischen t-Werts: Bei zweiseitigen Fragestellungen sprechen signifikante positive sowie negative t-Werte für die Alternativhypothese. Es ist darum nötig, auf jeder Seite der t-Verteilung eine Entscheidungsgrenze festzulegen. Damit insgesamt das gewünschte α-Niveau von 5% erreicht wird, darf der t_{krit} auf jeder Seite nur 2,5% der Fläche abschneiden.

> *Die Angabe der Auftretenswahrscheinlichkeit eines t-Werts bezieht sich bei **SPSS** automatisch auf eine zweiseitige Hypothese. Bei einem positiven empirischen t-Wert wird also nicht nur die Fläche berechnet, die dieser Wert nach rechts abschneidet, sondern gleichzeitig auch der Bereich, den derselbe negative t-Wert nach links abtrennt. Das bedeutet, dass der von SPSS ausgegebene zweiseitige p-Wert immer doppelt so groß ist wie der in der t-Tabelle verzeichnete einseitige:* $p_{zweiseitig} = 2 \cdot p_{einseitig}$
>
> *Ein empirischer t-Wert ist bei einer ungerichteten zweiseitigen Hypothese signifikant, wenn die von SPSS angegebene Auftretenswahrscheinlichkeit kleiner als 5% ist.*

3.2.2 Gerichtete Hypothesen

Aufgrund inhaltlicher Überlegungen kann in einigen Untersuchungen bereits die erwartete Richtung der Mittelwertsdifferenz spezifiziert werden. Es liegt dann eine einseitige Fragestellung vor. In einem solchen Fall umfasst die Alternativhypothese alle Differenzen in der vorhergesagten Richtung. Ist die vorhergesagte Differenz positiv, so nimmt die Nullhypothese an, dass die Differenz Null oder kleiner Null ist. In der mathematischen Schreibweise sieht das so aus:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Ein signifikantes Ergebnis bei einer gerichteten Fragestellung spricht als nur dann für eine Annahme der Alternativhypothese, wenn die Mittelwertsdifferenz in der vorhergesagten Richtung auftritt.

> *Beim Arbeiten mit **SPSS** muss die angegebene zweiseitige Auftretenswahrscheinlichkeit für gerichtete Fragestellungen halbiert werden.*

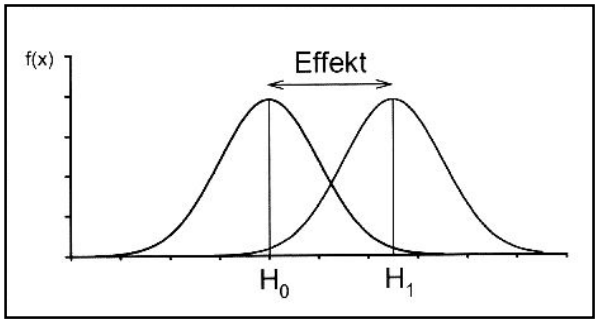
3.2.3 Vergleich von ein- und zweiseitigen Fragestellungen

Die Aufstellung gerichteter Hypothesen sollte also nur vorgenommen werden, sofern eine zugrunde liegende begründete und anerkannte Theorie es zulässt. Die ist allerdings bei genauer Betrachtung der Theorien nur selten gegeben. Häufig existieren Theorien für beide Richtungen der Mittelwertsdifferenz. In diesen Fällen wird daher zweiseitig getestet.

Die allgemeine Vorhersage eines Unterschieds in der mittleren Laufgeschwindigkeit zweier Sportlergruppen entspricht einer zweiseitigen Fragestellung. Für $n_1 = n_2 = 8$ und $\alpha = 0{,}05$ ergeben sich zwei kritische t-Werte: $t_{krit1(df=14)} = 2{,}145$ und $t_{krit2(df=14)} = -2{,}145$ (Tabelle B/t-Werte, Spalte für 0,975). Das Ergebnis $t_{(df=14)} = 1{,}9$ ist nicht signifikant ($p_{zweiseitig} < 0{,}05$). Ist die bessere Leistung der einen Gruppe bereits vorher bekannt oder wird sie auf der Basis einer Theorie vorhergesagt, so kann die Alternativhypothese auf die Differenz in der erwarteten Richtung eingeschränkt werden. Bei gleichen Bedingungen ergibt sich (Tabelle B/t-Werte, Spalte für 0,95):

$t_{krit(df=14)} = 1{,}761$, $t_{emp(df=14)} = 1{,}9$ ist signifikant ($p_{einseitig} < 0{,}05$).

3.2.5 Die nonzentrale Verteilung



Die Festlegung eines Populationseffektes spezifiziert die Alternativhypothese und greift eine bestimmte nonzentrale Verteilung heraus. Diese liegt soweit von Null entfernt, wie es der Effekt angibt. Auch die nonzentrale Verteilung kann einer empirisch gefundenen Mittelwertsdifferenz eine Auftretenswahrscheinlichkeit zuordnen, in diesem Fall also unter der Annahme eines Effekts einer bestimmten Größe.

Im Gegensatz zur Verteilung unter der Nullhypothese ist die nonzentrale Verteilung meist nicht symmetrisch. Die Form bzw. die Schiefe der Stichprobenkennwerteverteilung unter der Alternativhypothese ist über den so genannten Nonzentralitätsparameter λ (Lambda) bestimmbar. Dieser errechnet sich aus:

$$\lambda = \Phi^2 \cdot N$$

N ist die Anzahl aller Versuchspersonen, also $n_1 + n_2$; n ist die Anzahl der Versuchspersonen in einer Bedingung bzw. in einer „Zelle des Versuchsplans"; Φ^2 ist ein Effektstärkenmaß (siehe folgenden Abschnitt)

Diese Formel ist bei der Konstruktion eines t-Test von entscheidender Bedeutung.

3.3 Effektgrößen

Generell unterscheiden wir Effekte auf zwei Ebenen: Empirische Effekte, die das Ergebnis einer Untersuchung beschreiben, und Populationseffekte, die entweder angenommen oder aus den empirischen Daten geschätzt werden müssen. Die Größe eines empirischen Effekts ist für die inhaltliche Bewertung eines signifikanten Ergebnisses wichtig, da durch eine Erhöhung des Stichprobenumfangs theoretisch jeder noch so kleine Effekt signifikant gemacht werden kann.

Das Maß eines Effekts sollte standardisiert sein, um die Effekte verschiedener Untersuchungen miteinander vergleichen zu könne. Nur durch den Vergleich können in der praktischen Forschung Ergebnisse vernünftig interpretiert oder Theorien und Erklärungen weiterentwickelt und bestätigt werden. Es gibt zwei Möglichkeiten, ein standardisiertes Effektmaß mathematisch auszudrücken: als Distanz zwischen den Populationsmittelwerten oder als Varianzquotient.

3.3.1 Effekt als Distanz zwischen Populationsmittelwerten

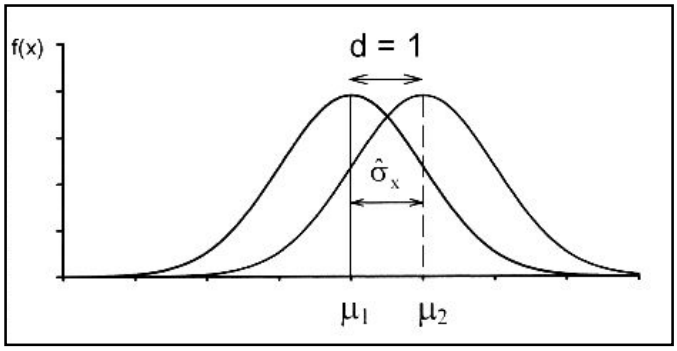
Ein absoluter Effekt (ein empirisch belegter Unterschied in den Ergebnissen) wird an der Streuung der Populationsverteilung innerhalb einer Stichprobe bzw. Bedingung standardisiert. Dabei wird angenommen, dass die Populationsstreuungen der zwei Populationen, aus denen die Stichproben stammen, theoretisch gleich groß sind (Varianzhomogenität). Die standardisierte Distanz kennzeichnet der Buchstabe d:

$$d = \frac{\mu_1 - \mu_2}{\sigma_x}$$

$\mu 1; \mu 2$: Mittelwerte der Populationen, aus denen die Stichproben gezogen werden

σ_x : Streuung der Population innerhalb einer Bedingung

Der empirische Effekt ergibt sich durch Einsetzen der Werte aus der Untersuchung. Dabei dient die Wurzel aus der gemittelten Varianz der beiden Stichprobenvarianzen („gepoolte“ Streuung) als bester Schätzer der Populationsstreuung:



Der Kennwert d ist wie ein Maß für eine Standardabweichung zu interpretieren. Er gibt an, um wie viele Streuungseinheiten sich zwei Gruppen unterscheiden, unabhängig von der zu Grunde liegenden Populationsstreuung in einer bestimmten Untersuchung.

d gestattet ausschließlich einen Vergleich zwischen zwei Gruppen, es bietet aber den großen Vorteil, dass empirische Effekte von unterschiedlichen Untersuchungen miteinander verglichen und bewertet werden können.

Folgende Konventionen sind nach Cohen (1988) für die Effektgröße d für einen t-Test mit unabhängigen Stichproben gültig:

- kleiner Effekt: d = 0,20
- mittlerer Effekt: d = 0,50
- großer Effekt: d = 0,80

3.3.2 Effektgrößen als Varianzquotient

Das Prinzip des Varianzvergleichs betrachtet nicht die Distanz oder den Abstand der Populationsmittelwerte voneinander, sondern bezieht sich auf den Abstand der Populationsmittelwerte von ihrem Mittelwert, dem so genannten Gesamtmittelwert. Diese entspricht der Varianzdefinition. Wenn Populationsmittelwerte voneinander signifikant verschieden sind, so existiert ein systematischer Effekt. Anders gesagt: die Populationsmittelwerte variieren systematisch um den Gesamtmittelwert. Die Varianz der Mittelwerte heißt deshalb systemische Varianz. Diesen Namen erhält sie, weil die Variation von Bedingungsmittelwerten (und der dahinter stehenden Populationsmittelwerte) in einer Untersuchung im Idealfall allein durch die kontrollierte (systematische) Manipulation seitens des Forschers zustande gekommen ist. Sie berechnet sich durch die Varianzformel, angewendet auf Populationsmittelwerte und für eine Anzahl von p Gruppen.

$$\sigma^2_{sys} = \frac{\sum_{i=1}^{p} (\mu_i - \bar{\mu})^2}{p}$$

Die Standardisierung der systematischen Varianz erfolgt nun an der Populationsvarianz innerhalb einer Bedingung. Diese Populationsvarianz trägt den Namen Residualvarianz. Dieses Verhältnis ist ein Maß für den Effekt, genannt Varianzquotient Φ^2 (Phi):

$$\Phi^2 = \frac{\sigma^2_{sys}}{\sigma^2_x}$$

Die Residualvarianz σ^2_x heißt unglücklicherweise häufig auch Fehlervarianz. In dieser Bezeichnung steckt die Überlegung, dass jede Abweichung der Werte von ihrem jeweiligen Populationsmittelwert einen Fehler zum Ausdruck bringt. Eine „fehlerfreie Messung“ ist also nur bei einer Populationsstreuung von Null möglich. Es ist besser, hier von der Residualvarianz („Restvarianz“) zu sprechen.

Φ^2 variiert von Null bis unendlich. Wenn die Residualvarianz bzw. die unsystematische Varianz gegen Null geht, läuft Φ^2 gegen unendlich.

Ein anschauliches Maß für den Effekt in der Population ist der Varianzquotient Ω^2 (Omega). Anstatt durch die Residualvarianz wird die systematische Varianz durch die Gesamtvarianz geteilt. Auf Populationsebene entspricht letztere im Fall des t-Test für unabhängige Stichproben der Summe der systematischen und der Residualvarianz.

$$\Omega^2 = \frac{\sigma^2_{sys}}{\sigma^2_{Gesamt}} = \frac{\sigma^2_{sys}}{\sigma^2_{sys} + \sigma^2_x}$$

Das Maß Ω^2 variiert nur zwischen 0 und 1, da der Zähler dieses Bruchs nie größer werden kann als der Nenner, und Varianzen aufgrund der Quadrierung immer positiv sind.

Ist die systematische Varianz $\sigma^2_{sys} = 0$, so ist auch der Effekt $\Omega^2 = 0$. Dieser Fall tritt dann auf, wenn die Populationsmittelwerte nicht variieren und kein systematischer Unterschied zwischen den Populationen besteht.

Ein Effekt von $\Omega^2 = 1$ kann nur dann auftreten, wenn es keine Residualvarianz gibt, also bei $\sigma^2_x = 0$. In diesem Fall sind Zähler und Nenner der Formel für Ω^2 identisch. Die Gesamtvarianz besteht demzufolge ausschließlich aus systematischer Varianz, und die Variation der Werte ist allein auf die experimentellen Unterschiede der beiden Gruppen zurückzuführen. Innerhalb einer Bedingung hätten alle Versuchspersonen exakt denselben Wert erzielt. Diese Situation tritt in den Sozialwissenschaften nie auf, da immer noch andere Faktoren als die experimentelle Manipulation die gemessenen Werte beeinflussen.

Theoretisch hätte die experimentelle Manipulation aber auf diese Weise die größte denkbare Auswirkung bzw. den größtmöglichen Effekt.

Insgesamt drückt das Effektstärkenmaß Ω^2 aus, wie groß der Anteil der systematischen Varianz an der Gesamtvarianz ist. Multipliziert man den Wert für Ω^2 mit 100, so lässt sich dieser Anteil in Prozent angeben. Ω^2 ist dann ein Maß dafür, wie viel Prozent der Gesamtvarianz durch die systematische Varianz aufgeklärt wird. Der „Rest" der Gesamtvarianz, die Residualvarianz, heißt deshalb auch unaufgeklärte Varianz.

Da die systematische Varianz durch die experimentelle Variation entstanden ist, lautet die Interpretation des Effekts Ω^2 in Bezug auf ein Experiment: Der Effekt Ω^2 gibt an, wie viel Prozent der Gesamtvarianz durch die Verschiedenheit der experimentellen Bedingungen auf Populationsebene aufgeklärt werden. Ω^2 fungiert damit als ein prozentuales Maß, das die Größe des Einflusses der experimentellen Manipulation anschaulich erfasst. Inhaltliche Überlegungen bei Planung oder Bewertung von Experimenten arbeiten deshalb in der Regeln mit Ω^2. Cohen hat auch hier wieder Werte zur Abstufung vorgeschlagen (es ist allerdings immer wichtig, die inhaltlichen Überlegungen im Auge zu behalten):

- kleiner Effekt: $\Omega^2 = 0{,}01$
- mittlerer Effekt: $\Omega^2 = 0{,}06$
- großer Effekt: $\Omega^2 = 0{,}14$

3.3.3 Schätzungen und Interpretation von Effektgrößen

$$f^2 = \frac{t^2 - 1}{N}, \quad f^2 = \hat{\Phi}^2$$

f^2 schätzt aus den empirischen Daten das Verhältnis von systematischer Varianz zu Residualvarianz. Es stellt eine Schätzung des Populationseffekts Φ^2 da und ist deshalb genau so wenig anschaulich wie letzteres. Auf Populationsebene bietet Ω^2 Abhilfe für dieses Problem, in dem es die systematische Varianz ins Verhältnis zur Gesamtvarianz setzt. Auch für Ω^2 gibt es einen Schätzer aus vorliegenden Daten, ω^2 (klein Omega Quadrat).

$$\omega^2 = \frac{f^2}{1 + f^2} \quad \text{bzw.} \quad f^2 = \frac{\omega^2}{1 + \omega^2}, \quad \omega^2 = \hat{\Omega}^2$$

Softwarehinweis: Das Programm GPower verwendet für die Berechnung der Teststärke eines t-Tests das Effektstärkenmaß d. Dies ist jedoch kein Problem, da sich das Maß ω^2 leicht in d überführen lässt. Die hier dargestellte Umrechnung gilt jedoch nur für den Vergleich von zwei Gruppen im Rahmen eines t-Tests für unabhängige Stichproben.

$$d = 2 \cdot f = 2 \cdot \sqrt{\frac{\omega^2}{1-\omega^2}}$$

3.3.5 Effektgrößen auf der Stichprobenebene

Das Effektstärkenmaß ω^2 nutzt die empirischen Daten, um einen Effekt auf der Populationsebene zu schätzen. Die Größe eines Effekts lässt sich aber auch auf der Ebene der Stichprobe bestimmen, ohne daraus Schlüsse für die Ebene der Population ziehen zu wollen. Diese Effektgröße auf Stichprobenebene ist nicht identisch mit dem Populationsschätzer ω^2.

Der Effektgröße η^2 (Eta-Quadrat) gibt den Anteil der aufgeklärten Varianz an der Gesamtvarianz auf der Stichprobenebene an und wird z.B. von dem Programm SPSS als Effektgröße verwendet. Die Berechung erfolgt über die Quadratsumme des systematischen Effekts, geteilt durch die gesamte Quadratsumme.

$$\eta^2 = \frac{QS_{sys}^2}{QS_{Gesamt}^2} = \frac{QS_{sys}^2}{QS_{sys}^2 + QS_x^2}$$

Die Berechung von η^2 ist ebenfalls mit Hilfe des t-Werts möglich. So wie f^2 als Schätzer der Effektgröße Φ^2 (auf Populationsebene) dient, wird dieses Maß mit dem Index S auch für Stichproben verwendet. f_s^2 gibt das Verhältnis der Quadratsumme des systematischen Effekts zu der Quadratsumme des Residuums an:

$$f_S^2 = \frac{QS_{sys}}{QS_x}$$

Um einen Effekt aus empirischen Daten zu bestimmen, ist die Umrechnung eines t-Werts zu den Stichprobeneffektgrößen f_s^2 und η^2 besonders wertvoll. Die Formeln lauten:

$$f_S^2 = \frac{t^2}{df} \qquad \eta^2 = \frac{f_s^2}{1+f_s^2}$$

Die Effektgröße η^2 gibt den Anteil der Varianzaufklärung auf der Ebene der Stichproben an, die Effektstärke Ω^2 auf der Ebene der Population.

3.4 Die Entscheidungsregel beim t-Test

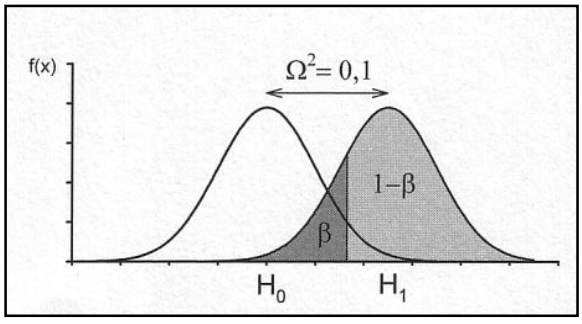
	In Wirklichkeit gilt die H_0	In Wirklichkeit gilt die H_1
Entscheidung zugunsten der H_0	Richtige Entscheidung	β-Fehler
Entscheidung zugunsten der H_1	α-Fehler	Richtige Entscheidung

Der α-Fehler (Fehler 1.Art) ist die Entscheidung für die H_1, obwohl in Wirklichkeit die H_0 gilt. Die α-Fehler-Wahrscheinlichkeit ist $\alpha = p(H_1 | H_0)$.
Der β -Fehler (Fehler 2.Art) ist die Entscheidung für die H_0, obwohl in Wirklichkeit die H_1 gilt. Die β -Fehler-Wahrscheinlichkeit ist $\beta = p(H_1 | H_0)$.

3.4.1 β –Fehler und Teststärke

Erst bei einer ausreichend kleinen β-Fehler-Wahrscheinlichkeit erlaubt also ein nicht signifikantes Ergebnis die Entscheidung für die Nullhypothese. Diese Wahrscheinlichkeit für den Fehler 2. Art sollte bei 10% oder weniger liegen. Ist sie größer so spricht ein nicht signifikantes Ergebnis für keine der beiden Hypothesen.
Es ist sehr wichtig, bereits vor Begin jeder Datenerhebung genaue Vorstellungen über die erwarteten Effekte zu haben, um so interpretierbare Ergebnisse am Ende der Untersuchung sicherzustellen.
Mit Hilfe des β-Fehlers kann auch eine Aussage darüber getroffen werden, wie gut ein t-Test konstruiert ist. Dies erfolgt durch eine Betrachtung der Teststärke oder Power eines t-Tests.
Die Teststärke ist die Wahrscheinlichkeit, die H_1 anzunehmen, wenn sie auch in Wirklichkeit gilt. Sie wird mit 1- β bezeichnet, da sie die Gegenwahrscheinlichkeit zu der β-Fehlerwahrscheinlichkeit ist.
Die Power eines t-Tests ist die Fähigkeit des Tests, einen Effekt zu finden, falls dieser tatsächlich existiert. Sie sollte mindestens 1- β = 0,9 betragen. Die Teststärke spielt bei der Planung und Beurteilung von t-Tests eine große Rolle. Ihre Bestimmung erfolgt entweder mit Hilfe des Nonzentralitätsparameters λ oder mit dem Computerprogramm GPower. Die Berechnung erfolgt nach folgender Formel:

$$\lambda = \Phi^2 * N$$



Die Berechnung der Teststärke eines bereits durchgeführten t-Tests ist dann notwendig, wenn ein nicht signifikantes Ergebnis auftritt und der Stichprobenumfang nicht im Vorfeld geplant wurde.

Wie schon das Signifikanzniveau α ist die Festlegung der akzeptierten Fehlerwahrscheinlichkeit β von inhaltlichen Überlegungen abhängig. Als Richtlinie schlagen wir eine Teststärke von 1- β = 0,9 vor. Die Wahrscheinlichkeit, die Alternativhypothese abzulehnen obwohl sie in Wirklichkeit gilt, liegt dann bei β = 0,1.

3.4.2 Die Determinanten des t-Tests

Die Größe des β-Fehlers und damit der Teststärke ist von drei Dingen abhängig: dem festgelegten Signifikanzniveau α, der Stichprobengröße und dem angenommenen Effekt. Zusammen mit dem β-Fehler bilden sie die vier Determinanten eines t-Tests.

Die Fehlerwahrscheinlichkeit

Nach der Bestimmung eines Signifikanzniveaus α ist auch die β-Fehler-Wahrscheinlichkeit festgelegt, solange die Stichprobengröße und der angenommene Effekt nicht verändert werden. Eine Verkleinerung von α bedingt eine Vergrößerung von β und damit eine Verkleinerung der Teststärke.

Der Einfluss des Effekts

Bei einem kleinen angenommenen Effekt liegen die Verteilungen der H_0 und der H_1 eng zusammen, sie überschneiden sich in der Regel stark (es sei denn, die Streuungen der Verteilungen sind extrem gering). Ein Signifikanzniveau α = 0,05 hat einen großen β-Fehler zur Folge, der t-Test hat eine geringe Teststärke. Bei einem größeren angenommenen Effekt wird die β-Fehler-Wahrscheinlichkeit bei gleichem α = 0,05 und gleichen Streuungen kleiner, die Teststärke größer.

Einfluss der Stichprobengröße

Je größer die Anzahl an Werten bzw. Versuchspersonen ist, desto schmaler werden die Stichprobenkennwerteverteilungen der H_0 und der H_1, ihre Streuungen werden kleiner. Das bedeutet, dass sie sich bei großen Stichproben und einem identischen angenommenen Effekt weniger überschneiden. Dies hat zur Folge, dass der kritische t-Wert bei konstantem Signifikanzniveau kleiner wird.

Zusätzlich nimmt der Stichprobenumfang Einfluss auf die Berechnung des t-Werts für die empirische Mittelwertsdifferenz. Mit zunehmender Stichprobengröße wird der Standfehler der Mittelwertsdifferenz kleiner. Dieser Standardfehler steht bei der Berechnung des t-Werts im Nenner, der t-Wert wird bei gleicher empirischer Mittelwertsdifferenz also größer. Ein

größerer t-Wert ist unter der Annahme der Nullhypothese unwahrscheinlicher und wird leichter signifikant. Je größer die Stichprobe, desto eher ein signifikantes Ergebnis.
Bei Stichproben größer als 30 schmiegt sich die t-Verteilung bereits eng an eine Normalverteilung an und die Wahrscheinlichkeiten für t-Werte verändern sich nur noch geringfügig.
Die Bestätigung eines vorhergesagten, kleinen Effekts in einem Experiment erfordert demzufolge eine größere Anzahl an Versuchspersonen, bei angenommenen großen Effekten reicht eine im Vergleich kleinere Anzahl an Beobachtungen aus.

3.4.3 Die Stichprobenumfangsplanung

Um einen eindeutigen interpretierbaren t-Test zu konstruieren, in denen der α- und β-Fehler hinreichend klein sind, dürfen die Stichprobengrößen nicht zu klein sein. Damit aber ein signifikantes Ergebnis nur auftritt, wenn ein inhaltlich bedeutsamer Effekt vorliegt, dürfen die Stichproben auch nicht zu groß sein. Man sagt, sie seien „optimal“. Dazu müssen sie vor Durchführung des Experiments berechnet werden. Eine solche Stichprobenumfangsplanung erfordert die Festlegung eines Signifikanzniveaus, der gewünschten Teststärke und eines inhaltlich bedeutsamen Effekts. Auch bestimmen inhaltliche Überlegungen die Wahl der drei Größen. Die TPF-Tabellen geben für die gewünschte Teststärke einen λ-Wert an.

$$N = \frac{\lambda_{\alpha;Teststärke}}{\Phi^2} = \frac{\lambda_{\alpha;Teststärke}}{\left(\frac{\Omega^2}{1-\Omega^2}\right)}$$

Am günstigsten ist, wenn die Anzahl pro Gruppe gleich ist, also:

$$n_1 = n_2 = \frac{N}{2}$$

3.4.4 Konfidenzintervall für eine Mittelwertsdifferenz

Da Mittelwertsdifferenzen t-verteilt sind, wird zur Konstruktion des Konfidenzintervalls anders als bei einfachen Mittelwerten die t-Verteilung unter $n_1 + n_2 - 2$ Freiheitsgraden genutzt.

untere Grenze: $(\bar{x}_1 - \bar{x}_2) - t_{(\alpha/2;df=n_1+n_2-2)} \cdot \hat{\sigma}_{\bar{x}_1-\bar{x}_2}$

obere Grenze: $(\bar{x}_1 - \bar{x}_2) + t_{(\alpha/2;df=n_1+n_2-2)} \cdot \hat{\sigma}_{\bar{x}_1-\bar{x}_2}$

$$\hat{\sigma}_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_{21}^2}{n_2}}$$

3.5.1 Der t-Test für abhängige Stichproben

Der t-Test für abhängige Stichproben betrachtet die Differenz der Werte jeder einzelnen Versuchsperson. Durch das Bilden der Differenz geht nur der Unterschied der Messwerte zwischen der ersten und der zweiten Messung in die Auswertung mit ein.

$d_1 = x_{i1} - x_{i2}$ i ist die Nummer der Vp, 1 oder 2 die jeweilige Bedingung

Der Stichprobenkennwert des t-Tests für abhängige Stichproben ist der Mittelwert der Differenzen aller erhobenen Versuchspersonen:

$$\bar{x}_d = \frac{\sum_{i=1}^{N} d_i}{N}$$

Da dieser t-Test die Verteilung der Mittelwerte von Differenzen betrachtet, ergibt sich eine andere Schätzung der Streuung der Stichprobenkennwerteverteilung als bei unabhängigen Stichproben (also ein anderer Standardfehler):

$\hat{\sigma}_{\bar{x}_d} = \frac{\hat{\sigma}_d}{\sqrt{N}}$ Wobei $\hat{\sigma}_d = \sqrt{\frac{\sum_{i=1}^{n} \left(d_i - \bar{x}_d\right)^2}{N-1}}$: geschätzte Populationsstreuung der Differenzen

Die Formel zur Berechnung des t-Werts ist ähnlich der für unabhängige Stichproben: An der Stelle zweier Differenzen von Mittelwerten stehen hier zwei Mittelwerte von Differenzen.

$$t_{\text{abhängig}} = \frac{\bar{x}_d - \mu_d}{\hat{\sigma}_{\bar{x}_d}}$$

Unter der Nullhypothese ist die Populationsmittelwertsdifferenz $\mu_d = 0$, deshalb vereinfacht sich die Formel zu:

$$t_{\text{abhängig}} = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{x}_d}}$$

Die Bewertung des t-Werts erfolgt analog zur Bewertung nach einem t-Test für unabhängige Stichproben. Die Freiheitsgradzahl berechnet sich aus df = N – 1 (Anzahl Messwertepaare bzw. Versuchspersonen minus Eins).

Die statistischen Hypothesen eines t-Tests für abhängige Stichproben lauten bei einer einseitigen bzw. zweiseitigen Fragestellung:

H_0: $\mu_d \leq 0$ bzw. $\mu_d = 0$

H_1: $\mu_d > 0$ bzw. $\mu_d \neq 0$

Berechnung der Effektstärke

Aufgrund der Komplexität wird sich hier auf das Effektstärkemaß auf Stichprobenebene bezogen. Das Effektstärkenmaß „partielles Eta-Quadrat“ (η_p^2) gibt den Anteil der aufgeklärten Varianz auf der Stichprobenebene an. Die Berechnung erfolgt mit Hilfe des t-Werts.

$$f^2_{S(abhängig)} = \frac{QS_{sys}}{QS_x} = \frac{t^2}{df} \qquad \eta_p^2 = \frac{QS_{sys}}{QS_{sys} + QS_x} = \frac{f_s^2}{1 + f_s^2}$$

Die ermittelte Effektgröße variiert bei abhängigen Stichproben nicht nur in Abhängigkeit vom Erfolg der experimentellen Manipulation, sondern auch als Funktion der Stärke der Abhängigkeit der Daten. Daher gibt es keine festen Konventionen für kleine, mittlere und große Effekte bei abhängigen Stichproben.

Teststärkeanalyse

Ein Maß für die Stärke der Abhängigkeit der Stichproben ist die Korrelation r (späteres Kapitel). Eine Berechnung der Teststärke erfolgt bei abhängigen Stichproben mittels folgender Formel:

$$\lambda_\alpha = \frac{2}{1-r} \cdot \phi^2_{unabhängig} \cdot N \qquad \text{mit} \qquad \phi^2_{unabhängig} = \frac{\Omega^2_{unabhängig}}{1 - \Omega^2_{unabhängig}}$$

Im Unterscheid zur unabhängigen Stichprobe geht hier die in der Untersuchung aufgetretene Abhängigkeit der Messwerte in Form der Korrelation in die Formel mit ein.

Um die Teststärke für einen empirisch ermittelten Effekt (aus vorliegenden abhängigen Daten) zu berechnen, gilt die nur leicht veränderte Formel für den Nonzentralitätsparameter λ.

$$\lambda = f^2_{S(abhängig)} * N$$

Stichprobenumfangsplanung

Die Stichprobenumfangsplanung ist identisch zu einem t-Test für unabhängige Stichproben mit der Ausnahme, dass sich eine halb so große Versuchspersonenzahl N ergibt, weil jede Versuchsperson zwei Werte abgibt.

$$N = \frac{\lambda_{\alpha;1-\beta}}{f^2_{abhängig}}$$

Für die Stichprobenumfangsplanung anhand der Konvention der Effektstärken für unabhängige Stichproben ist die Korrelation zwischen den wiederholten Messungen von

Bedeutung. Diese Korrelation ist leider in den wenigsten Fällen vor einer Studie bekannt. Eine konservative Möglichkeit ist es, keine Abhängigkeit zwischen den Messungen anzunehmen, also r = 0- Die Annahme errechnet in den meisten Fällen zu große Stichprobenumfänge. Eine Alternative ist die Annahme einer kleinen, aber substantiellen Korrelation zwischen den wiederholten Messzeitpunkten, z.B. eine Korrelation zwischen 0,2<r<0,4.

$$N = \frac{\lambda_{\alpha;1-\beta}}{\phi^2_{unabhängig}} \cdot \frac{(1-r)}{2}$$

3.5.2 Der t-Test für eine Stichprobe

Dieses Verfahren (one sample t-test) ist immer dann sinnvoll, wenn entweder der Populationsmittelwert bekannt ist (durch Normierungsstichproben) oder Annahmen über die Größe eines Populationsmittelwerts getestet werden soll.

Die zentrale Frage lautet hier: Wie groß ist die Wahrscheinlichkeit, dass die Stichprobe aus der Referenzpopulation stammt?

Bei einem vorliegenden Populationsmittelwert bedeutet ein signifikantes Ergebnis, dass die Stichprobe nicht zu der herangezogenen Referenzpopulation gehört oder zumindest eine eigene Subpopulation mit einem anderen Mittelwert bildet. Dieser t-Wert errechnet sich wie folgt:

$t_{df} = \frac{\bar{x} - \mu_0}{\hat{\sigma}_{\bar{x}}}$ mit df=N-1	(Mittelwert der Stichprobe, Populationsmittelwert der Referenzpopulation, Standardfehler des Mittelwerts)

Die Nullhypothese nimmt an, dass die Population, aus der die Stichprobe stammt, mit der Referenzpopulation identisch ist. Die Populationsmittelwerte wären dann identisch (Varianzhomogenität vorausgesetzt). Die Alternativhypothese dagegen postuliert einen Unterscheid zwischen der Population der Stichprobe und der Referenzpopulation.

Der t-Wert ist bei diesem Verfahren ein standardisiertes Maß für den Abstand des Stichprobenmittelwerts von dem betrachteten Populationsmittelwert. Ist die Wahrscheinlichkeit kleiner als das Signifikanzniveau, diesen t-Wert oder einen vom Betrag her größeren zu erhalten, so entstammt die Stichprobe wahrscheinlich nicht der angenommenen Population. Der Stichprobenmittelwert weicht signifikant von dem zu Grunde gelegten Populationsmittelwert ab.

3.6 Der Ablauf der Hypothesenprüfung mittels eines t-Tests

1 Aufstellen einer Hypothese

2 Prüfung der Vorraussetzungen

3 Festlegung eines Populationseffekts

4 Festlegung des Signifikanzniveaus

5 Stichprobenumfangsplanung

6 Bestimmung von t_{krit}

7 Prüfung des t_{emp} auf Signifikanz

8 Interpretation des Ergebnisses

1 Aufstellung einer Hypothese

Zunächst sollte eine möglichst spezifische Hypothese aufgestellt werden. Hieraus folgt dann die genaue Definition der Nullhypothese und einer Alternativhypothese. Durch die formale Schreibweise ist zu erkennen, ob es sich um eine ein- oder zweiseitige Fragestellung handelt.

2 Prüfung der Vorraussetzungen

Es stellen sich drei Fragen:

A: Sind die Daten intervallskaliert? Bei rang- oder nominalskalierten Daten ist die Anwendung des t-Tests nicht möglich.

B: Sind die Stichproben voneinander unabhängig oder abhängig?

C: Ist der Levene-Test der Varianzgleichheit signifikant? Wenn ja, dann ist eine Freiheitsgradkorrektur erforderlich (in SPSS automatisch ausgegeben). Bei einem nicht signifikanten Ergebnis des Levene-Test darf Varianzhomogenität angenommen werden.

Theoretisch wird auch die Prüfung der Vorraussetzung der Normalverteilung des untersuchten Merkmals in der Population gefordert, genügend große Stichproben und gleich große Gruppen vermeiden aber das Problem, als Faustregel sollte jede Bedingung mindestens 30 Versuchspersonen umfassen.

3 Festlegung eines Populationseffekts

Die Festlegung der Größe eines relevanten Populationseffekts ist von inhaltlichen Überlegungen abhängig. Dabei ist es besonders sinnvoll, sich an bereits vorliegenden Studien oder Metaanalysen zu einer vergleichbaren Fragestellung zu orientieren (Ω^2).

4 Festlegung des Signifikanzniveaus

Das Signifikanzniveau liegt per Konvention meistens bei $\alpha = 0{,}05$.

5 Stichprobenumfangsplanung

Die Berechnung des optimalen Stichprobenumfangs erfordert neben dem Effekt noch die Festlegung der gewünschten Teststärke. Die Größe der gewählten Teststärke ($1-\beta$) hängt von inhaltlichen Überlegungen ab. Die Berechnung des Stichprobenumfangs erfolgt mit Hilfe des Nonzentralitätsparameters λ.

6 Bestimmung von t_{krit}

Die Bestimmung des t_{krit} ist erst nach Durchführung einer Datenerhebung sinnvoll, weil erst jetzt wirklich sicher ist, wie viele Werte der Versuchspersonen in die Auswertung eingehen.

7 Prüfung des t_{emp} auf Signifikanz

Jetzt erfolgt die Berechnung der empirischen Mittelwertsdifferenz. Nach Schätzung der Populationsstreuung ergibt sich ein empirischer t-Wert unter der Annahme der Nullhypothese. Ist der empirische t-Wert größer als t_{krit}, so ist das Ergebnis signifikant. Ist t_{emp} kleiner als t_{krit}, hat der Unterschied keine statistische Bedeutsamkeit.

8 Interpretation des Ergebnisses

Ein signifikantes Ergebnis spricht für die Annahme der Alternativhypothese und für die Existenz des erwarteten Effekts in der Population. Die Wahrscheinlichkeit, dass diese Entscheidung falsch ist und in den Populationen kein Effekt der angenommenen Größe existiert, ist kleiner als 5% (bei einem *Signifikanzniveau* von $\alpha = 0{,}05$).

Bei einem nicht signifikanten Ergebnis erfolgt die Beibehaltung der Nullhypothese: Es gibt keinen Effekt in der erwarteten Größe zwischen den Populationen. Die Wahrscheinlichkeit, dass diese Entscheidung falsch ist und ein solcher Effekt doch existiert, beträgt 10% (bei einer festgelegten *Teststärke* von $1-\beta = 0{,}9$).

3.7 Ungeplante t-Tests

Bei ungeplanten Untersuchungen ergeben sich sowohl beim Auftreten eines signifikanten sowie eines nicht signifikanten Ergebnisses Probleme.

Um ein signifikantes Ergebnis wirklich beurteilen zu können, ist die Berechnung des empirischen Effekts aus den Daten erforderlich.

Eine Nichtsignifikanz macht die Situation noch schwieriger: Da die Teststärke und der β-Fehler unbekannt sind, schließt ein nicht signifikantes Ergebnis die Alternativhypothese nicht aus. Für diese Beurteilung ist eine Teststärkeberechnung notwendig. Eine nachträgliche Berechnung wird als a posteriori Berechnung bezeichnet (im Gegensatz zu a priori: vorherig). Dies geschieht über den Nonzentralitätsparameter λ, die zugehörige Teststärke ist von den Freiheitsgraden abhängig und in den TPF-Tabellen abzulesen.

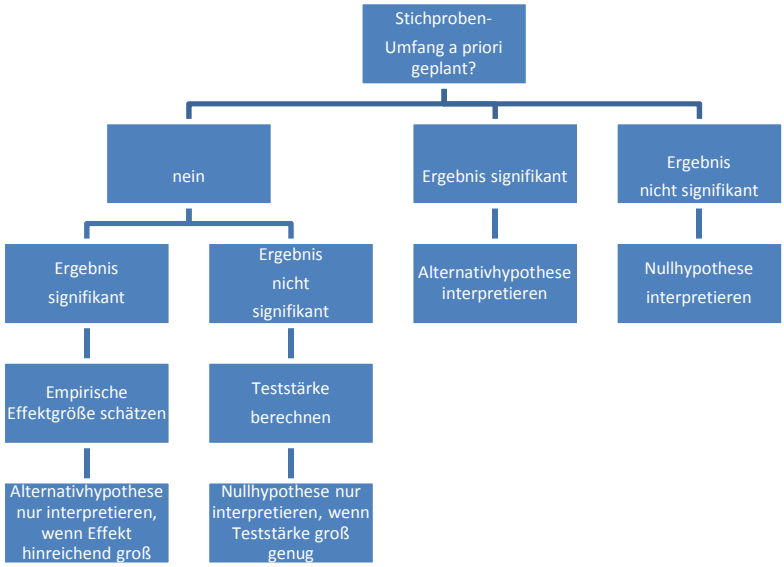
$\lambda = \Phi^2 * N$

Nichtsignifikante Ergebnisse sind nur bei ausreichend großer Teststärke bzw. ausreichend kleinem β-Fehler interpretierbar. Nur dann sprechen sie für die Nullhypothese. Außerdem schließt die Bestätigung der Nullhypothese nur Effekte einer bestimmten Größe mit der als Teststärke angegebenen Wahrscheinlichkeit aus. Über kleinere Effekte lassen sich keine Aussagen treffen.

Lesen eines t-Tests in der Literatur

Große Effekte können bei der Erforschung des menschlichen Verhaltens und Denkens nicht immer auftreten, weil in diesem Forschungsgebiet sehr viele Faktoren eine Rolle spielen. Jeder Faktor für sich macht da oft nur einen kleinen, aber durchaus nicht zu vernachlässigenden Effekt aus. Die Interpretation der Größe eines Effekts ist deshalb stark von dem Inhalt der Untersuchung abhängig.

Grundsätzlich gilt, dass nie eine Argumentation auf der Nichtsignifikanz eines Ergebnisses aufgebaut werden darf, es sei denn, die Teststärke ist ausreichend hoch oder der β-Fehler ist a priori auf einem inhaltlich sinnvollen Niveau festgelegt worden.



4. **Verfahren für Rangdaten**

Dieses Kapitel stellt drei statistische Auswertungsverfahren vor, die sich auf die Analyse von Daten auf Ordinalskalenniveau beziehen.

Durch eine Zuordnung von Rangplätzen wird eine künstliche Äquidistanz zwischen den Werten erzeugt, die viele mathematische Operationen wie z.B. die Mittelwertsbildung erst ermöglicht. Die nichtparametrischen Verfahren für Rangdaten arbeiten nicht mit Populationsparametern und –verteilungen, sondern legen ihre eigene Verteilung (die der Rangplätze) zu Grunde.

4.1 Der Mann-Whitney U-Test

Der U-Test für unabhängige Stichproben oder auch Mann-Whitney U-Test ist ein Verfahren zur Auswertung eines Zwei-Gruppen-Experiments, dessen Bedingungen sich in einer unabhängigen Variable unterscheiden. Ähnlich wie der t-Test für unabhängige Stichproben prüft der U-Test, ob die Unterschiede in den zwei Gruppen bezüglich einer abhängigen Variable zufälligen oder systematischen Einflüssen unterliegen. Anders als der t-Test aber analysiert der U-Test die Messwerte nicht direkt, sondern die ihnen zugeordneten Rangplätze.

Der U-Test stellt primär ein Instrument für ordinalskalierte Daten dar. Darüber hinaus ist er aufgrund seiner größeren Vorraussetzungsfreiheit in folgenden Fällen dem t-Test für unabhängige Stichproben vorzuziehen:

- Die Intervallskalenqualität der abhängigen Variable ist zweifelhaft.
- Das Merkmal folgt in der Population keiner Normalverteilung.
- Die Annahme der Varianzhomogenität ist verletzt, so dass der t-Test für unabhängige Stichproben keine zuverlässigen Ergebnisse mehr liefert.

Der U-Test prüft die Nullhypothese, dass kein Unterschied zwischen den beiden untersuchten Gruppen hinsichtlich des erhobenen Merkmals besteht.

4.1.1 Zuordnung der Rangplätze

Für die Berechnung des U-Werts ist es zuerst notwendig, alle Messwerte in eine gemeinsame Rangreihe zu bringen. Danach wird für jede Gruppe die Summe der Rangplätze sowie der durchschnittliche Rangplatz der Gruppe berechnet.

$$T_i = \sum_{m=1}^{n_i} R_{mi}$$

i : Gruppe

n_i : Anzahl der Versuchspersonen in Gruppe i

R_{mi} : Rangplatz der m-ten Versuchsperson in der Gruppe i

Zur Kontrolle der Berechnung beider Summen dient folgende Beziehung: Die beiden Rangsummen T_1 und T_2 müssen zusammen die Gesamtsumme der Rangplätze ergeben. Da den Messwerten die Zahlen 1 bis N zugeordnet worden sind, lässt sich die Gesamtsumme der Zahlen 1 bis N über folgende Formel berechnen:

$$\sum_{m=1}^{N_i} R_m = \frac{N \cdot (N+1)}{2} = T_1 + T_2$$

R_m : Rangplatz der m-ten Versuchsperson

N : Anzahl Beobachtungen ($N=n_1+n_2$)

Die mittlere Rangsumme einer Gruppe ergibt sich aus der Division dieser Summe durch die Anzahl der Versuchspersonen in einer Gruppe:

$$\overline{R_i} = \frac{T_i}{n_i}$$

T_i : Summe der Rangplätze in Gruppe i

n_i : Anzahl der Versuchspersonen in Gruppe i

4.1.2 Der U-Wert und U'-Wert

Rangplatzüberschreitungen

Der statistische Kennwert U wird gebildet, indem für jeden Rangplatz einer Person der einen Gruppe die Anzahl an Personen aus der anderen Gruppe gezählt wird, die diesen Rangplatz überschreitet.

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - T_1$$

n_1 : Anzahl Untersuchungseinheiten in Gruppe 1

n_2 : Anzahl Untersuchungseinheiten in Gruppe 2

T_1 : Rangsumme für Gruppe 1

Rangplatzunterschreitungen

Diese Summe nennt sich U'.

$$U' = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2+1)}{2} - T_2$$

4.1.3 Signifikanzprüfung beim U-Test

Wie auch bei den parametrischen Verfahren erfolgt die Signifikanzprüfung des U-Tests unter Annahme der Nullhypothese.

Die Nullhypothese

Die Beziehung zwischen U und U' zeigt folgende Formel:

$$U = n_1 \cdot n_2 - U'$$

Ein größerer Unterschied zwischen den Gruppen führt in jedem Fall zu einem größeren Unterschied zwischen U und U'. Besteht kein Unterschied zwischen den beiden Gruppen, gibt es genauso viele Rangunter- wie Rangüberschreitungen.

Die Nullhypothese des U-Tests lautet deshalb:

$$U = U'$$

Unter der Annahme der Nullhypothese (U=U') lässt sich aus dem Zusammenhang von U und U' eine erwarteter mittlerer U-Wert μ_U berechnen:

$$\mu_U = \frac{n_1 \cdot n_2}{2}$$

Weicht der empirische U-Wert sehr stark in positive oder negative Richtung von μ_U ab, so spricht das gegen die Interpretation der Nullhypothese. Die bei wiederholter Ziehung von Stichproben resultierende Verteilung um den Mittelwert μ_U ist symmetrisch.

Streuung der Stichprobenkennwerteverteilung

Die Streuung der U-Werte um den Mittelwert μ_U kann bei unverbundenen Rängen durch folgende Formel errechnet werden:

$$\sigma_U = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$$

Signifikanzprüfung bei großen Stichproben

Wenn eine der Stichproben n_1 oder n_2 größer als 20 ist und sich n_1 und n_2 nicht zu stark unterscheiden, nähert sich die Kennwerteverteilung der U-Werte einer Normalverteilung an. Das ermöglicht es, die Standardnormalverteilung als Prüfverteilung heranzuziehen. Der empirische U-Wert, der Mittelwert μ_U, sowie die Streuung der U-Verteilung σ_U werden einfach in die bekannte Formel zur Berechnung eines z-Werts eingesetzt:

$$z = \frac{x_i - \mu}{\sigma_x}$$

Bezogen auf den U-Test ergibt sich die Formel für die Berechnung des z-Werts zu:

$$z_U = \frac{U - \mu_U}{\sigma_U}$$

4.1.4 Verbundene Ränge

Versuchspersonen mit gleichen Messwerten gehen zunächst normal in die Zuordnung mit ein, bei N Versuchspersonen ist der letzte zugeordnete Rangplatz also immer noch die Zahl N. Allerdings wird aus den zugeordneten Rangplätzen für gleiche Messwerte ein mittlerer Rangplatz gebildet.

Die Streuung der Stichprobenkennwerteverteilung ändert sich im Fall von verbundenen Rängen. Die Formel zu deren Berechnung aus dem vorangehenden Abschnitten wird wie folgt korrigiert:

$$\sigma_{U_{corr}} = \sqrt{\frac{n_1 \cdot n_2}{N \cdot (N-1)}} \cdot \sqrt{\left(\frac{N^3 - N}{12} - \sum_{i=1}^{k} \frac{t_i^3 - t_i}{12} \right)}$$

N : n_1+n_2

T_i : Anzahl der Personen, die sich Rangplatz i teilen

k : Anzahl der verbundenen Ränge

Die Korrekturformel für die Streuung der U-Verteilung sieht für jeden der k Fälle, in welchem ein verbundener Rangplatz vorliegt, eine Adjustierung vor, wobei die Anzahl t_i der Personen berücksichtigt wird, die sich den Rangplatz i teilen.

4.1.5 Teststärke und Stichprobenumfangsplanung

Der Vorteil der nichtparametrischen Verfahren besteht darin, dass sie ohne mathematische Annahmen über die Verteilung des untersuchten Merkmals auskommen. Sie sind verteilungsfreie Verfahren. Allerdings resultieren hieraus auch Nachteile, es fehlen eigene Prozeduren zur Berechnung von Effekt- und Teststärken.

Die Stichprobenumfangsplanung

Die Stichprobenumfangsplanung beim U-Test erfolgt mit Hilfe des bereits bekannten Vorgehens beim t-Test.

$$N_{(t-Test)} = \frac{\lambda_{df=1;\alpha;1-\beta}}{\Phi} = \frac{\lambda_{df=1;\alpha;1-\beta}}{\frac{\Omega^2}{1-\Omega^2}}$$

Teststärkeanalyse

Die Teststärkebestimmung beim U-Test erfolgt wie schon die Stichprobenumfangsplanung mit Hilfe des t-Tests. Auch hier dienen die Konventionen des t-Tests zur groben Orientierung.

$$\lambda_{df=1;\alpha} = \Phi^2 \cdot N_{(t-Test)}$$

4.2 Der Wilcoxon-Test

Der Wilcoxon-Test oder W-Test ist das nichtparametrische Pendant zum t-Test für abhängige Stichproben. Er ist immer in solchen Fällen das nichtparametrische Verfahren der Wahl, wenn die Messwerte zweier Stichproben in irgendeiner Weise voneinander abhängig sind.

Der W-Test für abhängige Stichproben arbeitet ebenfalls mit der Analyse einer Rangreihe. Die Bildung der für die statistische Auswertung notwenigen Rangplätze und der dadurch erzeugten künstlichen Äquidistanz erfolgt in vier Schritten:

1. Zuerst wird die Differenz der Messwertpaare gebildet, indem der Wert der zweiten Bedingung oder Gruppe von dem der ersten abgezogen wird. Dieser Schritt entspricht der Vorgehensweise beim t-Test für abhängige Stichproben.
2. Von jeder Differenz wird der Betrag gebildet, das Vorzeichen als ignoriert. Die Differenzen der Größe Null werden ignoriert und gehen nicht in die Auswertung mit ein. Die Anzahl N der Rangplätze wird um die Anzahl der Nulldifferenzen reduziert.
3. Die Absolutbeträge der Differenzen bilden einer Rangreihe. Der kleinste Differenzbetrag erhält den Rang 1, der nächste den Rang 2 us. Liegen zwei oder mehr vom Betrag her gleiche Differenzen vor, so entspricht die Zuordnung der Rangplätze zu verbundenen Rängen der Methode des U-Tests für unabhängige Stichproben.
4. Nach der Zuordnung wird jeder absolute Rangplatz, der zu einer negativen Differenz gehört, mit einem negativen Vorzeichen versehen.
Die auf diese Weise gekennzeichneten Rangplätze heißen „gerichtete Ränge“ (signed ranks). Sie bilden die Grundlage für die statistische Auswertung des Wilcoxon Tests. Unter der Annahme der Nullhypothese gibt es keinen Unterschied zwischen den zwei verglichenen Messzeitpunkten. Die Differenzen sollten zufällig zu Stande gekommen sein. Unter Annahme der Nullhypothese sollte die gleiche Anzahl an positiven wie negativen Differenzen auftreten, die ebenso von ihren Beträgen her ungefähr gleich sein sollten. Die Alternativhypothese behauptet das Gegenteil. Wenn z.B. die erreichten Messwerte in der zweiten Messung wesentlich größer ausfallen als in der ersten, sollten mehr und größere positive Differenzen und weniger und kleinere negative Differenzen auftreten. Die Umkehrung dieser Aussage ist ebenfalls eine Alternativhypothese. Zur statistischen Bewertung werden die Summen der gerichteten Rangplätze gebildet. Dr vom Betrag her kleinere Wert ist der Testwert W des Wilcoxon-Tests.

$$W = \left| \min\left(\sum R_{positiv}, \sum R_{negativ}\right) \right|$$

Ein empirischer W-Wert muss gleich dem kritischen oder kleiner als ein kritischer W-Wert sein, damit das Ergebnis signifikant ist. Dieser kritische Wert ist von der Anzahl der Versuchspersonen abhängig, deren Differenzen nicht Null ergeben.

4.3 Der Kruskal-Wallis H-Test

Der Kruskal-Wallis H-Test ist ein Verfahren für die statistische Auswertung ordinalskalierter Daten von mehr als zwei unabhängigen Gruppen. Er bietet eine Alternative für die einfaktorielle Varianzanalyse ohne Messwiederholung, wenn deren mathematische Vorraussetzungen nicht erfüllt sind. Deshalb heißt er auch Rangvarianzanalyse.

Der H-Test arbeitet wie die besprochenen U- und W-Tests mit zugewiesenen Rangreihen. Wie die Varianzanalyse testet er die Nullhypothese, dass die zu Grunde liegenden Verteilungen der untersuchten Gruppen identisch sind. Der Hintergrund des Tests besteht in der Überlegung, dass die Rangplätze bei Zutreffen der Nullhypothese zufällig über alle Gruppen verteilt sein müssten. Die Prüfung erfolgt analog zur Varianzanalyse unspezifisch: Die Alternativhypothese besagt lediglich, dass sich mindestens eine der Gruppen von den anderen unterscheidet.

Die Zuordnung der Ränge zu den Messwerten erfolgt analog zum U-Test für unabhängige Stichproben: Allen Messwerten wird unabhängig von ihrer Gruppenzugehörigkeit ja nach Größe des Messwerts eine ganze Zahl zwischen 1 und N zugeordnet. Bei gleichen Messwerten wird ein mittlerer Rang aus den zugehörigen Rängen gebildet. Darauf folgt in weiterer Analogie zum U-Test die Bestimmung der Summe T_i der Ränge (R_m) in jeder Gruppe.

$$T_i = \sum_{m=1}^{N} R_{im}$$

Anschleißend wird die Rangsumme jeder Gruppe quadriert und durch die Anzahl der Versuchspersonen in der entsprechenden Gruppe geteilt. Die resultierenden Werte aller Gruppen werden addiert. Für p Gruppen ergibt das die folgende Berechnung:

$$\sum_{i=1}^{p} \frac{T_i^2}{n_i} = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + ... + \frac{T_p^2}{n_p}$$

Für die Berechnung des Kennwerts H geht dieser Wert zusammen mit der Gesamtzahl aller Versuchspersonen N in folgende Formel ein:

$$H = \left[\frac{12}{N \cdot (N+1)} \right] \cdot \left[\sum_{i=1}^{p} \frac{T_i^2}{n_i} \right] - 3 \cdot (N+1)$$

Die Verteilung des H-Werts nähert sich einer χ^2-Verteilung mit df = p – 1 Freiheitsgraden an, wenn die Versuchspersonenanzahl in keiner der Gruppen kleiner als n_i = 5 ist. Die Freiheitsgrade ergeben sich aus der Anzahl der Gruppen minus Eins. Bei ausreichend großen Gruppen kann also der H-Wert mit einem χ^2-Wert verglichen werden. Überschreitet der H-Wert den kritischen χ^2-Wert, so ist das Ergebnis signifikant.

Stichprobenumfangsplanung und Teststärke

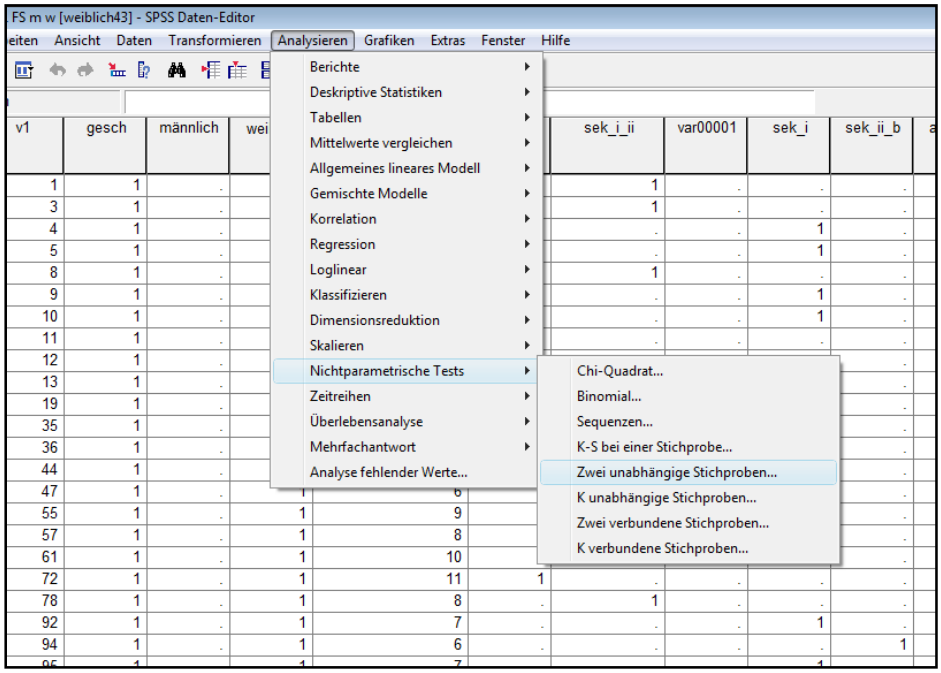
Für die Stichprobenumfangsplanung und Teststärkeberechnung in einem Kruskal-Wallis H-Test wird vorgeschlagen, sich an den Berechnungen einer entsprechenden einfaktoriellen Varianzanalyse ohne Messwiederholung zu orientieren.

Anhang SPSS

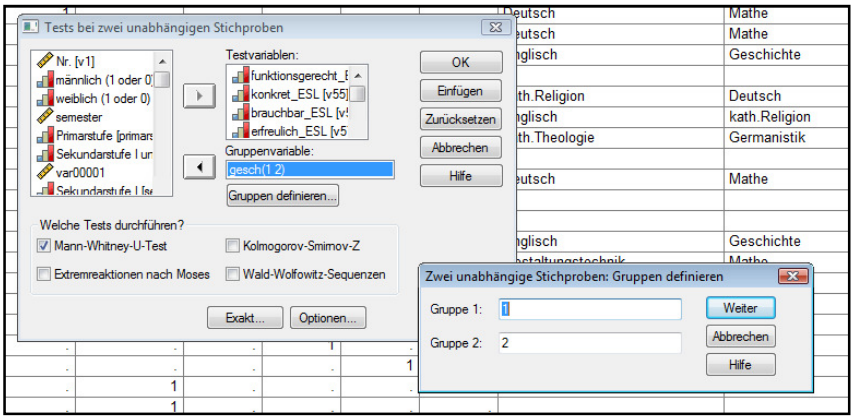
Der Anhang zu SPSS gibt noch einmal die wichtigsten Schritt des Workshops graphisch wieder.

1. Mann-Whitney-U-Test

1.1 Analysieren Nichtparametrische Tests Zwei unabhängige Stichproben

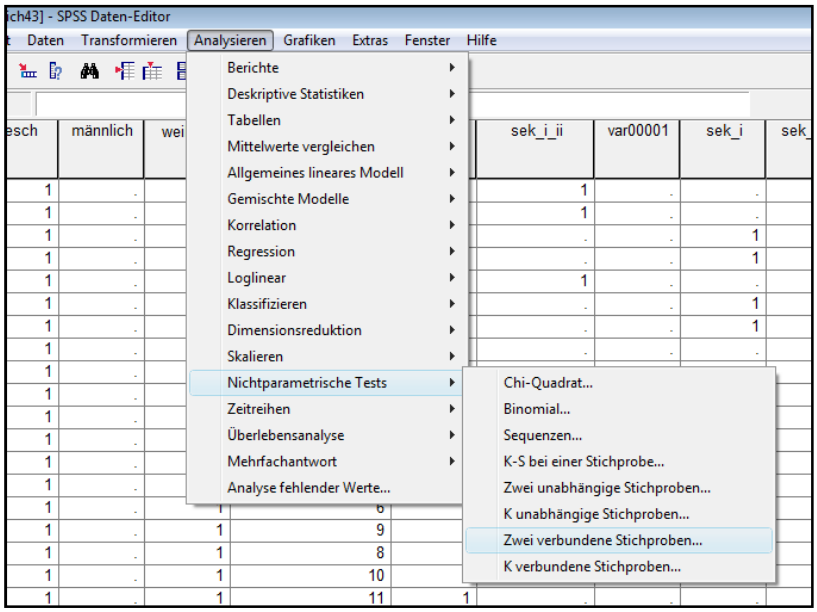


1.2 Gruppenvariable aussuchen Gruppen definieren Testvariable(n) angeben

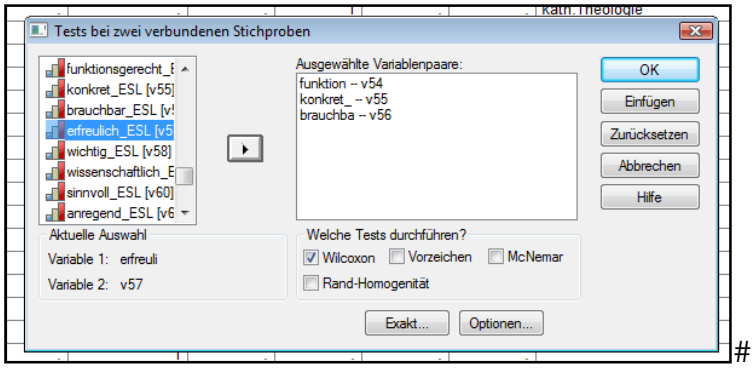


2. Der Wilcoxon-Test

2.1 Analysieren Nichtparametrische Tests Zwei verbundene Stichproben



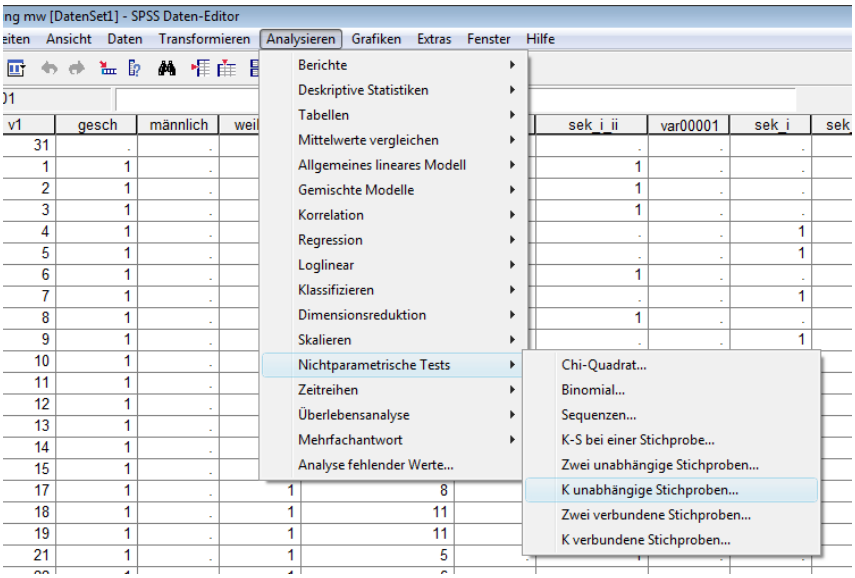
2.2 Variable 1 anklicken Variable 2 anklicken Pfeil benutzen, um dieses Variablenpaar dem Bereich „Ausgewählte Variablenpaare“ hinzuzufügen (Wilcoxon sollte schon automatisch aktiviert sein)



#

3. Der Kruskal-Wallis H-Test

3.1 Analysieren Nichtparametrische Tests K unabhängige Stichproben



3.2 Gruppenvariable aussuchen Gruppenbereich definieren Testvariable(n) angeben