Glossar für einige wichtige statistische Begriffe

Abhängige bzw. unabhängige Variablen. Je nach Theorie die einer Untersuchung zugrunde liegt, gibt es Vermutungen über die Kausalrichtung in der Variablen zueinander stehen. Eine Variable kann dann als Ursache und eine andere als Wirkung aufgefasst werden. Die Wirkung wird als *abhängige Variable* bezeichnet, da sie von der Ursache abhängig ist. Die Ursache wird hingegen als *unabhängige Variable* bezeichnet, wenn im Rahmen der Theorie und/ oder der Untersuchung keine zusätzlichen Variablen berücksichtigt werden, die die Ursache beeinflussen. Bei vielen Studien steht eine einzige zentrale Größe als abhängige Variable im Vordergrund. Darauf können mehrere unabhängige Variablen einen Einfluss haben. So wird die körperliche Gesundheit (abhängige Variable) bestimmt von verschiedenen unabhängigen Variablen (Sport, Ernährung, Trinkverhalten, Gesundheitsverhalten, Alter, Geschlecht, genetischen Prädispositionen etc.). Als *Kontrollvariablen* bezeichnet man solche unabhängigen Variablen, die einen Einfluss haben könnten und daher statistisch berücksichtigt werden müssen, aber nicht zentrales Forschungsinteresse der Studie sind.

Abhängige Daten. Bei Interventionsstudien sind die Ergebnisse nach der Intervention abhängig von den Merkmalsausprägungen vor der Intervention. Damit eine Veränderung sichtbar wird müssen verschiedene Messzeitpunkte miteinander verglichen werden. Die Messwerte stehen dadurch in einer zeitlichen Abhängigkeit. Für abhängige Daten sind besondere statistische Testverfahren vorgeschlagen worden. Für einige Auswertungen kann es zudem sinnvoll sein den Unterschied zwischen den Zeitpunkten zu berechnen. Dieses sog. *Delta* lässt sich dann statistisch als eine Größe für die Wirkung der Intervention verwenden.

Abweichungsmaße, Streuungsmaße. Mittelwerte und andere ↗ *Maße der zentralen Tendenz* geben zwar einen Eindruck über die Daten insgesamt, die einzelnen Messwerte weichen jedoch in der Regel auch von der zentralen Tendenz ab. Diese Abweichungen werden durch Abweichungsmaße erfasst. ↗ *Standardabweichung, Streuung oder Varianz* zeigen wie intervallskallierte Messwerte (↗ Messung) vom arithmetischen Mittel abweichen. Abweichungen vom Median werden als Interquartilsabstand angegeben. Dabei wird die Messwerteverteilung nach der Größe sortiert, die ersten 25% der Daten sind das erste Quartil, die 50%-Grenze ist der Median und die ersten 75% der Daten sind das dritte Quartil. Der Interquartilsabstand ist der Bereich vom ersten zum dritten Quartil. 50% aller Daten liegen innerhalb dieser Grenzen. Bei nominalen Daten kann der Modalwert den häufigsten Wert angeben. Eine Abweichung ist hier durch Angabe der Häufigkeit und der Prozentzahl ersichtlich. Auch kann es hier sinnvoll sein, den seltensten Wert ebenfalls zu bestimmen.

Alpha-Fehler, Beta-Fehler, *Test Power*. Ein Signifikanztest (↗ statistische Signifikanz) befindet den Unterschied zwischen zwei Kennwerten (z.B. Mittelwerten) dann als signifikant, wenn der Unterschied so groß ist, dass es nach den Gesetzen der Wahrscheinlichkeitsrechnung nur eine geringe Wahrscheinlichkeit (↗ P-Wert) dafür gibt, dass *kein Unterschied* besteht. Wie gering sollte dafür die Wahrscheinlichkeit sein? Es handelt sich um eine Übereinkunft, dass üblicherweise bei einer Wahrscheinlichkeit von 5% (und darunter) von ↗ Signifikanz gesprochen wird. Das bedeutet dann, dass ein Signifikanztest, der zwei Kennwerte mit einer Wahrscheinlichkeit von 5% für ähnlich hält, zu dem Schluss kommt, dass eine Signifikanz vorliegt. Die Annahme, dass die Kennwerte ähnlich sind, wird daher verworfen und die dazu passende Alternativhypothese (↗ Hypothesenarten) wird akzeptiert. Die Signifikanzgrenze von 5% gibt dabei gleichzeitig den Fehler der Entscheidung über eine Signifikanz an. Denn mit einer 5%igen Wahrscheinlichkeit ist die Nullhypothese ja korrekt. Wenn sie dennoch verworfen wird – und damit Signifikanz angenommen wird – ist dies mit einer Wahrscheinlichkeit von 5% falsch. Wenn man aufgrund eines Signifikanztests davon ausgeht, dass eine Signifikanz vorliegt, macht man mit eben jener Wahrscheinlichkeit die als Signifikanzgrenze festgelegt wird einen Fehler. Dieser Fehler wird *Alpha-Fehler* genannt. Mitunter wird auch nur von Alpha gesprochen. Dieses Alpha ist nicht zu verwechseln mit *Cronbachs Alpha* einem Maß für die internen Konsistenz einer Fragebogenskala. Der Alpha-Fehler ist der Fehler fälschlicherweise von Signifikanz auszugehen obwohl keine Vorliegt. Demgegenüber steht ein *Beta-Fehler* (Fehler 2. Art) dieser gibt an wie hoch die Wahrscheinlichkeit dafür ist fälschlicherweise davon auszugehen, dass keine Signifikanz vorliegt. Hier wird also eine vorliegende Signifikanz übersehen. Den Alpha-Fehler kann man für eine Untersuchung frei wählen. So ist Alpha = 5% eine übliche Signifikanzgrenze. Diese kann auch strenger gewählt werden, z.B. mit 1% oder 0,1%. Der Beta-Fehler hängt aber von Umständen ab, die nicht beeinflusst werden können. Der Beta-Fehler kann nicht direkt gewählt und festgelegt werden. Der Beta-Fehler ist gegeben durch die Empfindlichkeit des eingesetzten statistischen Testverfahrens. Testverfahren mit einem hohen Beta-Fehler übersehen Signifikanzen. Eins minus Beta wird als *Test-Power* bezeichnet. Das ist die Wahrscheinlichkeit mit der ein Test eine vorliegende Signifikanz auch findet.

Alpha-Fehler-Adjustierung. In der Regel sind Signifikanztests in der Lage, nur zwei Kennwerte (z.B. Mittelwerte) miteinander zu vergleichen. Einige Fragestellungen bzw. Hypothesen machen mehrere Vergleiche zwischen jeweils zwei Kennwerten nötig, um die Hypothese insgesamt prüfen zu können. Beantworten z.B. drei Personengruppen einen Fragebogen (Gruppe A, B, C), so kommt man auf insgesamt drei paarweise Vergleiche (A mit B; A mit C und B mit C). Allgemein gilt: Anzahl der Vergleiche = [Anzahl der Gruppen mal [Anzahl der Gruppen minus Eins]] geteilt durch Zwei. So ergeben sich für vier Gruppen bereits: (4 x 3)/2 = 6 Vergleiche. Wenn die Hypothese relativ offen formuliert ist und generell nach Unterschieden zwischen den Gruppen gefragt wird, so wächst die Wahrscheinlichkeit, einen Unterschied zu finden, je mehr Vergleiche möglich werden. Da man ja bei jedem Paarvergleich einen ↗ Alpha-Fehler von 5% begeht, summieren sich die Fehler von Paarvergleich zu Paarvergleich. Bei drei Vergleichen macht man also einen viel höheren Fehler als bei nur einem. Höhere Fehler als 5% sind jedoch nach der oben angesprochenen Vereinbarung nicht signifikant. Um insgesamt nur auf einen Fehler von 5% zu kommen, müssen für jeden Einzelvergleich strengere Alpha-Fehler-Grenzwerte festgelegt werden. Für 3 Vergleiche ergibt sich z.B. ein Wert von 1,7%, bei vier Vergleichen sind es 1,3%, bei 10 Vergleichen 0,5%, usw. Eine solche Anpassung der Signifikanzgrenze für mehrfaches Testen heißt Alpha-Fehler-Adjustierung. Eine Alternative für die Berechnung vieler Signifikanztests, die nur jeweils zwei Kennwerte vergleichen, ist die Varianzanalyse (↗ Varianzanalyse, ANOVA) oder die multiple Regression.

Chi-Quadrat-Test. Der Chi-Quadrat-Test ermöglicht den Vergleich von erwarteten Häufigkeiten mit tatsächlich beobachteten Häufigkeiten. Erwartet man aufgrund von Vorerfahrungen oder aus der Literatur zum Beispiel, dass jeder vierte männliche Österreicher Raucher ist, so würde man bei 100 befragten Personen 25 Raucher erwarten. Der Chi-Quadrat-Test vergleicht die erwarteten 25 Raucher dann mit den tatsächlich im Rahmen einer Befragung vorgefundenen Rauchern. Im Rahmen eines Chi-Quadrat-Tests können beliebig viele verschiedene Häufigkeiten miteinander verglichen werden. So ergibt sich beim Chi-Quadrat-Test auf eine Gleichverteilung die erwartete Häufigkeit als Mittelwert der beobachteten Häufigkeiten. Aufgrund geringer Voraussetzungen kann der Chi-Quadrat-Test immer berechnet werden, wenn es um Häufigkeiten geht und eine bestimmte oder mehrere bestimmte Häufigkeiten erwartet werden können. Der Chi-Quadrat-Test ermittelt einen Chi-Quadrat-Wert, für den zusammen mit den sog. Freiheitsgraden (in der Regel Zahl der Messwerte minus eins) die Wahrscheinlichkeit bekannt ist. Die Wahrscheinlichkeit ist das Ergebnis des Tests. Man spricht von einer ↗ statistischen Signifikanz, wenn diese Wahrscheinlichkeit kleiner als der vorher festgelegte ↗ Alpha-Fehler ist. Da der Chi-Quadrat-Test für ganz verschiedene Hypothesen über die Verteilung von Häufigkeiten angewendet werden kann gibt es zahlreiche sehr unterschiedliche Chi-Quadrat-Tests. Hier ist darauf zu achten den konkreten Zweck des Tests genau zu prüfen und ihn im Rahmen einer wissenschaftlichen Arbeit nicht einfach nur als Chi-Quadrat-Test zu bezeichnen sondern z.B. als Chi-Quadrat-Test für Kreuztabellen (4-Felder-Tabellen) oder auf Gleichverteilung oder auf Normalverteilung. Der Test arbeitet mit einer Näherungsgleichung, die bei keinen Stichproben auch mal daneben liegen kann. Der Chi-Quadrat-Test wird daher heute viel seltener verwendet als noch vor einigen Jahren.

Fishers exakter Test. Ein besonders *sicherer* Test ist Fishers exakter Test, da er kaum an Voraussetzungen gebunden ist und immer berechnet werden kann, wenn es um den Vergleich zweier Prozentzahlen geht. Er ist damit eine exakte und bessere Alternative für den Chi-Quadrat-Test für Kreuztabellen (4-Felder-Tabellen). Eine Berechnung durch einen Computer ist aber rechenintensiv und kommt an Grenzen, wenn die Stichprobengröße ca. den Wert 1000 erreicht. Neben der exakten Variante dieses Tests gibt es für große Stichproben daher auch Näherungsformeln über den ↗ T-Test, die jedoch mit Vorsicht zu genießen sind. Fishers exakter Test liefert die Wahrscheinlichkeit für die Übereinstimmung zweier Prozentzahlen. Die Wahrscheinlichkeit ist das Ergebnis des Tests. Man spricht von einer ↗ statistischen Signifikanz, wenn diese Wahrscheinlichkeit kleiner als der vorher festgelegte ↗ Alpha-Fehler ist.

Hypothesenarten. Hypothesen werden aus Theorien bzw. konkreten theoretischen Annahmen und bereits publizierten wissenschaftlichen Studien logisch abgeleitet. Sie formulieren auf Grundlage dieser theoretischen Annahmen eine Vorhersage für den konkreten Fall der geplanten Untersuchung. Die Hypothese, die aus den theoretischen Grundlagen folgt heißt „*Alternativhypothese*“. Statistisch geprüft wird die Verneinung der Alternativhypothese. Diese wird als „*Nullhypothese*“ bezeichnet. Ist die Nullhypothese unwahrscheinlich, wird von statistischer ↗ Signifikanz gesprochen. Eine unwahrscheinliche Nullhypothese wird verworfen und die Alternativhypothese wird akzeptiert. Hypothesen die statistisch geprüft werden sollen, sollten idealerweise direkt so formuliert werden, dass der statistische Zugang in der Hypothese bereits deutlich wird. Die Statistik kennt *Unter-*

schiedshypothesen – bei denen Unterschiede zwischen zweien oder mehreren Untersuchungsgruppen (↗ Kategorien) – vermutet werden, und sie kennt *Zusammenhangshypothesen* – bei denen Zusammenhänge zwischen zweien oder mehreren Variablen vermutet werden. Hypothesen über *keinen* Unterschied und *keinen* Zusammenhang können nicht auf Signifikanz geprüft werden und sollten daher vermieden werden. Hypothesen können *einseitig* oder *zweiseitig* formuliert werden (↗ P-Wert). Im einseitigen Fall wird die Richtung des Unterschieds (z.B. Hypothese: Das Gehalt der Männer ist höher als das der Frauen.) oder des Zusammenhangs (z.B. Hypothese: Es besteht ein positiver Zusammenhang zwischen Alkohol und Krebsrisiko.) in der Hypothese konkret genannt. Ob diese Form der Hypothese gewählt wird, hängt von der Theorie bzw. der Studienlage ab. Ist diese uneindeutig, wird zweiseitig formuliert und offen gelassen, welche Kategorie höhere Werte zeigt (z.B. Hypothese: Die Schulklassen unterscheiden sich in ihren Leistungen.) oder welcher Zusammenhang vorliegt (z.B. Hypothese: Es besteht ein Zusammenhang zwischen Musikgenuss und Nervosität.).

Kategoriale Daten. In Unterschiedshypothesen werden Unterschiede in einer ↗ abhängigen Variable für verschiedene Untersuchungsgruppen – oder allgemeiner: Kategorien – angenommen. So wird z.B. ein Unterschied im Gehalt von Frauen und Männern, die beiden Kategorien Frauen und Männer als unabhängige Variable heranziehen und innerhalb der Kategorien das Gehalt feststellen. Eine Kategorie umfasst in der Regel mehrere Untersuchungseinheiten, also nicht nur einen Mann. Kategorien sind entweder durch die Untersuchung direkt gegeben (z.B. das Geschlecht) oder müssen nach der Untersuchung gebildet werden (z.B. Altersgruppen). Nominale Daten bilden automatisch Kategorien. Ordinale Daten oder Intervalldaten können nachträglich in Kategorien unterteilt werden (↗ Messung, Messniveau, Skalenniveau).

Korrelationen. Eine Korrelation beschreibt den statistischen Zusammenhang zwischen zwei Merkmalen. Beide Merkmale müssen in unterschiedlichen Ausprägungen vorkommen können. Ist das nicht der Fall, so kann keine Korrelation berechnet werden. Wird z.B. die Frage danach gestellt, ob die Zahl der Geburten und die Zahl der Störche einen Zusammenhang (also eine Korrelation) aufweist, so muss sowohl die Zahl der Störche, als auch die Zahl der Geburten variieren können. Es bietet sich hier an, die Zahl der Geburten und die Zahl der Störche pro Monat zu erheben. Dadurch erhält man Zahlenpaare aus Geburtenzahl und Storchenpopulation für jeden Monat. Es stellt sich nun die Frage, ob sich die Zahl der Störche und die Zahl der Geburten über das Jahr hinweg in die gleiche Richtung entwickelt, also ob mit ansteigender Zahl der Geburten auch die Zahl der Störche wächst und ob mit sinkender Zahl der Geburten auch die Zahl der Störche abnimmt. Ist es so, dass die Zahl der Störche und die Zahl der Geburten sich jeweils in die gleiche Richtung entwickeln, so spricht man von einer positiven Korrelation. Steigt jedoch die Zahl der Geburten, immer wenn die Zahl der Störche abnimmt (und umgekehrt: die Zahl der Geburten sinkt und gleichzeitig nimmt die Zahl der Störche zu), so spricht man von einer negativen Korrelation. Korrelationen können Zahlenwerte zwischen –1 und +1 annehmen. Dabei zeigt das Vorzeichen an, ob es sich um einen positive oder um eine negative Korrelation handelt. Je näher die Zahlenwerte bei 1 (bzw. –1) liegen, desto „perfekter“ ist der Zusammenhang. Ist der Zahlenwert jedoch 0, dann liegt gar keine Korrelation – also auch kein Zusammenhang – vor. Viele Zusammenhänge, die z.B. in der Psychologie beschrieben werden, haben relativ kleine Werte um 0,3 (bzw. –0,3), wohingegen z.B. in der Physik nicht selten Korrelationen um 0,9 (bzw. –0,9) gefunden werden können. Die Höhe einer Korrelation zu interpretieren ist daher nicht leicht. Es gibt zwar allgemein akzeptierte Einteilungen aber im konkreten Anwendungsfall können auch andere Grenzen für eine ausreichende oder nicht ausreichende Korrelation sinnvoll sein. Allgemein gilt eine Korrelation ohne Berücksichtigung des Vorzeichens ab 0,1 als klein, ab 0,3 als mittel und ab 0,5 als groß (Cohen, 1992). Ob eine Korrelation nicht eventuell doch auf das Fehlen einer Korrelation (Null-Korrelation) hinweist, kann nur durch einen Signifikanztest (↗ Statistische Signifikanz) entschieden werden. Erst, wenn eine Korrelation sich als signifikant herausstellt, kann sie interpretiert werden. Ist sie nicht signifikant, so kann man nicht davon ausgehen, dass ein Zusammenhang beobachtet wurde. Ist sie jedoch signifikant, so bedeutet das noch nicht, dass der beobachtete Zusammenhang kausal zu interpretieren ist. Es gibt vielleicht Studien, die zeigen, dass die Zahl der Störche mit der Zahl der Geburten in einigen Gegenden im Verlauf des Jahres tatsächlich korreliert. Das würde jedoch nicht bedeuten, dass die Störche die Kinder bringen. Es gibt verschiedene statistische Verfahren die die Korrelation in Abhängigkeit von der ↗ Messung, dem Messniveau, dem Skalenniveau berechnen. In der Regel wird dabei von linearen Zusammenhängen ausgegangen. Liegen tatsächlich nichtlineare Zusammenhänge vor, kann die Korrelationsberechnung fälschlicherweise auf eine Null-Korrelation verweisen.

x / y	Intervall	Dichotom (2-stufig: z.B. ja/nein)	Ordinal
Intervall	Produkt-Moment-Korrelation (Pearson)	Punktbiseriale Korrelation (Alternativ: T-Test) *Bei 1/0-Kodierung der dichotomen Variable ist die Produkt-Moment-Korrelation identisch mit der Punktbiserialen Korrelation.*	Rangkorrelation (Spearman) *Bei Kodierung der Ordinalskala mit 1, 2, 3, … ist der Wert mit der Produkt-Moment-Korrelation identisch*
Dichotom (2-stufig: z.B. ja/nein)		Phi-Koeffizient (über Chi-Quadrat). *Bei 1/0-Kodierung der dichotomen Variablen ist die Produkt-Moment-Korrelation identisch mit Phi.*	Biseriale Rangkorrelation (Alternativ: U-Test)
Ordinal			Rangkorrelation *Bei Kodierung der Ordinalskalen mit 1, 2, 3, … ist der Wert mit der Produkt-Moment-Korrelation identisch*

Mann-Whitney-U-Test. Besteht der Verdacht, dass die Voraussetzungen für einen ↗ T-Test verletzt sein könnten, kann am besten der U-Test von Mann und Withney berechnet werden.

Messung, Messniveau, Skalenniveau. Bei einer Messung werden empirische Gegebenheiten mit Zahlen versehen. Das Ziel ist dabei die Unterschiede, Ähnlichkeiten oder Relationen, in denen die empirischen Gegebenheiten zueinander stehen, mit den Zahlen bestmöglich wiederzugeben. Nach der Messung liegen nur mehr die Zahlen vor und es muss mitgeteilt werden und bekannt sein, wie die Messzuordnung erfolgte und was man aus den Zahlen ablesen darf und was aufgrund der Messung nicht interpretiert werden kann. *Nominalskala*: Die Zahlen werden ein-eindeutig den Objekten zugeordnet. Die Zahlen können die Objekte identifizieren. Die Höhe der Zahlen hat keinerlei Bedeutung. Beispiel: Zahlencode für Berufe, Bäcker:in = 234, Professor:in = 543, … *Ordinalskala*: Die Anordnung der Zahlen gemäß ihrer Größe entsprich einer Ordnung der empirischen Gegebenheiten. Diese wird aber nur grob wiedergegeben oder ist tatsächlich nur grob vorhanden. So kann der Abstand der Zahlen zueinander nicht als Abstand der empirischen Gegebenheiten zueinander interpretiert werden. Beispiel: höchster Bildungsabschluss: Pflichtschule = 1, Abitur = 2, Studium = 3, ... *Intervallskala*: Die Abstände zwischen den Zahlen können sinnvoll interpretiert werden. Zahlenverhältnisse können nicht sinnvoll interpretiert werden. Beispiel: Alter gemessen in Jahren. Wenn eine Person 2 Jahre älter ist als eine andere, wird das so bleiben, auch wenn Zeit vergeht. Wenn eine Person *exakt doppelt so alt* ist wie eine andere, ist das am nächsten Tag oder in der nächsten Stunde oder Minute schon nicht mehr korrekt. Das Zahlenverhältnis ist also nicht vernünftig interpretierbar. *Verhältnisskala*: Zahlenverhältnisse sind sinnvoll interpretierbar. Beispiel: Gehalt. Eine Verhältnisskala erfordert einen inhaltlich klaren und unveränderbaren Nullpunkt. Das ist beim Gehalt gegeben. Die exakte Zahlengröße ist bei dieser Skala nach der Messung immer noch veränderbar und nicht als unveränderbare Größe interpretierbar (z.B. kann das Gehalt in verschiedenen Währungen angegeben werden). *Absolutskala*: Bei einer Absolutskala ist eine nachträgliche Umrechnung der Zahlen in andere Maßeinheiten unsinnig. Z.B. ist die Zahl von Personen in einem Raum eine Zahl, die exakt diese Anzahl angibt und sinnvoll nicht mehr verändert werden sollte. Je nach Skalenniveau sind also verschiedene Eigenschaften interpretierbar und daher passende statistische Verfahren zu wählen.

Skalenniveau	**zentrale Tendenz**	**Abweichungsmaß**	**Anmerkung**
Nominal	**Modalwert (häufigster Wert)**	**Prozent, seltenster Wert**	
Ordinal	**Median** (Modalwert)	**(Inter)-Quartilsabstand** (Prozent, seltenster Wert)	
Intervall	**Mittelwert** (Median) (Modalwert)	**Standardabweichung, Varianz (Stichprobe)** (Inter)-Quartilsabstand (Prozent, seltenster Wert)	Median und Inter-(Quartilsabstand) sind mitunter besser, weil Ausreißer wenig ins Gewicht fallen.
Verhältnis	Wie Intervall, aber auch geometrisches Mittel (z.B. Zinsen)	wie Intervall	

Maße der zentralen Tendenz, (Mittelwert, Median, Modalwert). Messwerte einer Stichprobe unterscheiden sich in der Regel. *Maße der zentralen Tendenz* werden eingesetzt, um mit einer einzigen Zahl einen Eindruck über die Daten zu vermitteln. Je nach Skalenniveau (↗ Messung) kann ein arithmetischer Mittelwert (Intervallskalenniveau), der Median (Ordinalskala) oder der Modalwert (Nominalskalenniveau) benutzt werden. Der Mittelwert, als Summe aller Messwerte, geteilt durch die Anzahl der Messwerte liegt exakt in der Mitte der Messwerteverteilung. Er berücksichtigt dabei die Abstände zwischen den Messwerten. So ist der Mittelwert empfindlich gegenüber extremen Zahlenwerten, auch dann, wenn diese nur selten in der Stichprobe vorkommen. Der Median weist eine solche Empfindlichkeit nicht auf. Der Median ist die Mitte der nach der Größe sortierten Messwerte. Er teilt die Daten in zwei Hälften, so dass 50% der Messwerte kleiner als der Median sind und 50% darüber liegen. Da der Median die Abstände zwischen den Messwerten nicht berücksichtigt, kann er auch für ordinale Daten benutzt werden. Demgegenüber kann bei nominalen Daten nur der Modalwert herangezogen werden. Das ist der Messwert, der insgesamt am häufigsten vorkommt. Bei einigen Fragestellungen ergibt es sich, dass Mittelwert, Median und Modalwert exakt den gleichen Wert aufweisen. Dies ist jedoch nicht immer der Fall. Aus der Anordnung der drei Werte kann man Informationen über die Verteilung der Messwerte in der Stichprobe gewinnen. Bei Merkmalen, die durch extreme Antworten verzerrt sein könnten, ist der Median eventuell eine gute Wahl, auch dann, wenn die Daten Intervallskalenniveau aufweisen. Maße der zentralen Tendenz geben einen Eindruck über die Daten, die allerdings von dieser Tendenz in der Regel auch abweichen. Diese Abweichungen werden durch ↗ Abweichungsmaße erfasst.

P-Wert. Das Ergebnis eines Signifikanztests (↗ statistische Signifikanz) ist die Wahrscheinlichkeit dafür, dass die Nullhypothese – also das Gegenteil der eigentlich in der Hypothese formulierten Aussage – zutrifft (↗ Hypothesenarten). Es wird also die Wahrscheinlichkeit dafür bestimmt, dass der vermutete Unterschied nicht besteht bzw. der vermutete Zusammenhang nicht vorliegt. Da Wahrscheinlichkeit auf Englisch *Probability* heißt, wird sie mit dem Buchstaben „p" abgekürzt. p kann jedoch grundsätzlich auf zwei verschiedene Arten berechnet werden. p kann 1-*seitig* oder auch 2-*seitig* bestimmt werden. Welche der beiden Berechnungen im Einzelfall anzugeben ist, entscheidet sich durch die Hypothese, die mit dem Signifikanztest beantwortet werden soll. Eine zweiseitige ↗ Hypothese prüft, ob ein Unterschied besteht, ohne genauere Vermutungen darüber anzustellen, in welche Richtung der Unterschied weisen könnte. Eine einseitige Fragestellung geht darüber hinaus. Sie prüft nicht nur, ob allgemein ein Unterschied besteht, sondern zudem, ob er in die erwartete Richtung geht. Der 2-seitige Wert wird also bei ungerichteten Signifikanztests angegeben. Er ist immer exakt doppelt so hoch wie der entsprechende 1-seitige Wert. Der 1-seitige Wert hat es damit „leichter" signifikant zu werden, erfordert aber die genauere Hypothese. Der P-Wert gibt also die Wahrscheinlichkeit für die Nullhypothese an. Daraus lässt sich nicht errechnen, wie hoch die Wahrscheinlichkeit für die Alternativhypothese ist. Ist der P-Wert klein, wird die Nullhypothese verworfen. Die Alternativhypothese wird als Alternative zur Nullhypothese in diesem Fall akzeptiert. Sie ist dadurch weder bewiesen noch kann man davon ausgehen, dass Eins minus dem P-Wert die Wahrscheinlichkeit für die Alternativhypothese darstellt. Es kann nämlich sehr viele Alternativhypothesen geben. Die Wahrscheinlichkeit für die Alternativhypothese kann aus logischen Gründen niemals bestimmt werden. Der P-Wert für die Nullhypothese wird in einigen Statistikprogrammen als „Signifikanz" bezeichnet. Das ist genau genommen nicht korrekt, weil die Signifikanz eben erst vorliegt, wenn der P-Wert besonders klein ist.

Regressionsanalyse. Eine Regressionsanalyse prüft den Einfluss einer oder mehrerer unabhängiger Variablen auf eine einzigen abhängige Variable. Das Messniveau der unabhängigen Variablen und die Art der Kodierung ist zwar nicht beliebig, aber die Regressionsanalyse kann verschiedene Skalenniveaus für die unabhängigen Variablen leicht verarbeiten. So können unterschiedliche Variablen gleichzeitig berücksichtigt werden. Nominale oder ordinale Variablen werden z.B. als dichotome Dummy-Variablen mit den Zahlen 1 und 0 kodiert. Intervallskalierte unabhängige Variablen können direkt genutzt werden. Interaktionen zwischen Variablen lassen sich ebenfalls berücksichtigen indem die interargierenden Variablen zuvor miteinander multipliziert werden und dieses Produkt als neue unabhängige Variable zusätzlich berücksichtigt wird. Die Regressionsanalyse ist daher insgesamt sehr flexibel und vielseitig einsetzbar. Sie hat inzwischen viele klassische Tests wie einfache T-Tests abgelöst. Eine Regressionsanalyse liefert zwei verschiedene Informationen. (1) Ihr Ergebnis ist eine Gleichung mit der bei gegebenen unabhängigen Variablen, die Höhe der abhängigen Variable geschätzt werden kann. In diesem Sinne „lernt" eine Regressionsanalyse an gegebenen Daten wie der Zusammenhang zwischen den unabhängigen Variablen und der abhängigen ist. Danach kann dann die ermittelte Regressionsgleichung benutzt werden, um aus den unabhängigen Variablen eine noch unbekannte abhängige zu berechnen. So könnte aus Studien eine Gleichung für das Krebsrisiko erstellt worden sein, in die leicht abfragbare Größen wie Rauchen, Alkoholkonsum, Größe, Gewicht, Alter, Geschlecht etc. eingehen. Der Einsatz von Regressionsgleichungen zur Vorhersage von abhängigen Variable erfordert eine besondere Prüfung der Regressionsmodelle. Wie gut ein Modell insgesamt funktioniert kann z.B. durch eine Gesamtkorrelation R angegeben werden. Sie gibt an wie hoch die Vorhersage mit den wahren Werten der abhängigen Variable korreliert. Ist R hoch in der Nähe von Eins, dann ist das Modell perfekt. Diese Perfektion ist aber nicht immer das Ziel einer Regressionsanalyse. (2) Diese kann auch benutzt werden, um den Zusammenhang zwischen den unabhängigen Variablen und der abhängigen Variable statistisch zu prüfen. Hier steht nicht die Vorhersage im Vordergrund sondern der statistisch Test. Der Vorteil der Regressionsanalyse ist, dass gleichzeitig alle unabhängigen Variablen geprüft werden und in ihrem Zusammenspiel berücksichtigt werden. Wenn z.B. in einer Stichprobe die befragten Frauen durchschnittlich älter waren als die befragten Männer könnte es sein, dass ein mit dem T-Test ermittelter Unterschied im Gehalt nicht auf das Geschlecht sondern auf das Alter und die Berufserfahrung zurückgeführt werden könnte. Der T-Test sieht diesen Zusammenhang nicht. In der Regressionsanalyse, mit beiden Variablen, Alter und Geschlecht, zeigt dann z.B. dass das Alter eine Signifkanz und das Geschlecht keine mehr. Da die Regressionsanalyse gleichzeitig mehrere unabhängige Variablen enthalten kann, ist das Ergebnis nicht immer leicht zu interpretieren. Die Signifikanz einer Variable kann nur im Zusammenhang mit den ebenfalls berücksichtigten weiteren Variablen sinnvoll verstanden werden. Scheinbar paradoxe Effekte zeigen sich z.B. wenn Variablen signifikant werden, die genau besehen überhaupt nicht direkt mit der abhängigen Variable korrelieren. Z.B. hat Kaffee-Konsum keinen Einfluss auf den IQ aber wenn der Kaffee in die Regressionsgleichung aufgenommen wird, wird er signifikant. Kaffee als Grund für eine übersteigerte Nervosität beim IQ-Test verfälscht vielleicht die Ergebnisse und wird daher signifikant. Eine solche Variable heißt Supressor-Variable. Es gibt verschiedene Typen von Regressionsmodellen. Diese müssen für unterschiedliche abhängige Variablen gewählt werden. Ist diese intervallskaliert wird die klassische multiple Regression durchgeführt, für dichotome abhängige Variablen (ja/nein) wird die binär-logistische Regression herangezogen, für die Zeit bis ein Ereignis eintritt die Cox-Regression. Nichtlineare Zusammenhänge können durch vorhergehende mathematische Transformationen der unabhängigen Variablen geprüft werden. Die theoretische Planung einer Regressionsanalyse ist mitunter recht aufwendig. So sollen ja mehre Variablen gleichzeitig berücksichtigt werden. Es sind daher auch Verfahren vorgeschlagen worden, die aus einem Pool von Variablen Schritt für Schritt die signifikanten Variablen automatisiert herauswählen oder nicht signifikante Schrittweise ausschließen.

Standardabweichung, Streuung, Varianz. Die Standardabweichung, auch Streuung genannt, ist – vereinfacht gesprochen – ein Wert für die mittlere Abweichung der Messwerte vom Mittelwert (ohne Berücksichtigung der Abweichungsrichtung). Die Standardabweichung bzw. Streuung gibt damit einen Eindruck von der Variationsbreite der Antworten und damit zum Teil auch über die Messgenauigkeit. Bei ideal normalverteilen Messwerten liegt der ↗ Mittelwert zusammen mit dem ↗ Median und dem ↗ Modalwert exakt in der Mitte der Messwerteverteilung. Insgesamt 68% aller Antworten befinden sich dann in dem Messwertebereich zwischen dem Mittelwert minus der Streuung und dem Mittelwert plus der Streuung. Beispiel: Ein IQ-Test weist in der Regel einen Mittelwert von 100 und eine Streuung von 10 auf. Damit liegen 68% aller Menschen mit ihrem IQ zwischen einem IQ von 90 und 110. Die Varianz ist nichts anderes als das Quadrat der Streuung bzw. Standardabweichung.

Statistische Signifikanz. (Statistische Bedeutsamkeit) Jeder im Rahmen einer Messung gewonnene Messwert ist mit einer gewissen Fehlertoleranz behaftet. Die Ergebnisse einer Befragung sind daher nie exakt. Die Genauigkeit einer Messung kann in vielen Fällen mit Hilfe der Wahrscheinlichkeitsrechnung angegeben werden. In diesem Sinne bezeichnet z.B. die ↗ Streuung die Schwankungsbreite der Messwerte um den Mittelwert. Wenn nun zwei Kennwerte verglichen werden sollen, z.B. der Mittelwert des Gehalts für Männer mit dem Mittelwert des Gehalts für Frauen, so muss immer auch mitbedacht werden, dass beide Messwerte ungenau sind. Ein sichtbarer Unterschied in den Mittelwerten bedeutet nicht automatisch, dass sich die beiden Untersuchungsgruppen unterscheiden. Denn dieser Unterschied könnte auf Messfehler und natürliche Schwankungen innerhalb der Untersuchungsgruppen zurückzuführen sein. Ein statistischer Signifikanztest beantwortet die Frage, ob ein Unterschied zwischen zwei Kennwerten (z.B. Mittelwerten) durch Messungenauigkeiten erklärt werden kann. Erst wenn die Wahrscheinlichkeit dafür, dass kein Unterschied vorliegt, gering ist und unter der vorher festgelegten Signifikanzgrenze (in der Regel 5%, ↗ Alpha-Fehler) liegt, sagt man, dass die Unterschiede statistisch signifikant sind. D.h., dass ein statistischer Signifikanztest niemals behaupten würde, dass ein Unterschied zwischen den Kennwerten für Männer und den Frauen tatsächlich besteht. Statistisch signifikant heißt nur, dass es unwahrscheinlich (aber nicht unmöglich) ist, dass kein Unterschied besteht. Je nach erhobenen Daten müssen verschiedene Verfahren für die Signifikanzprüfung angewandt werden. Wichtige Testverfahren sind z.B.: ↗ T-Test, ↗ Fishers exakter Test, ↗ Chi-Quadrat-Test, ↗ Mann-Withney-U-Test, ↗ Varianzanalyse, Signifikanzprüfung einer ↗ Korrelation. Das wichtigste Ergebnis eines Testes ist die Wahrscheinlichkeit (↗ P-Wert) dafür, dass sich die Kennwerte nicht unterscheiden. Diese Wahrscheinlichkeit wird mit einem vorher festgelegten Grenzwert, der Signifikanzgrenze (↗ Alpha-Fehler) verglichen.

T-Test. Ein besonders gebräuchlicher Signifikanztest für den Vergleich von zwei Mittelwerten ist der T-Test (↗ statistische Signifikanz). Der T-Test besitzt jedoch einige Voraussetzungen, die erfüllt sein müssen, damit er berechnet werden kann. Diese Voraussetzungen sind allerdings nicht immer erfüllt. Zu den Grundvoraussetzungen gehört u.a., dass mit gutem Gewissen ein ↗ Mittelwert und die dazu gehörige ↗ Streuung berechnet werden können. Die Verteilung der Mittelwerte muss einer Normal- bzw. T-Verteilung folgen, was bei kleinen Stichproben Probleme machen kann. Bei Stichproben mit einer Gruppengröße von mindestens 25 bis 50 Personen pro Untersuchungsgruppe, liegt automatisch eine Normalverteilung der Mittelwerte vor (↗ zentraler Grenzwertsatz), so dass dann kein Problem bei der Anwendung des T-Tests besteht. Der T-Test berechnet einen t-Wert, für den zusammen mit den sog. Freiheitsgraden (in der Regel: Zahl der Messwerte minus eins) die Wahrscheinlichkeit bekannt ist. Die Wahrscheinlichkeit ist das Ergebnis des Tests. Man spricht von einer ↗ statistischen Signifikanz, wenn diese Wahrscheinlichkeit kleiner als der vorher festgelegte ↗ Alpha-Fehler ist.

Validität, prognostische. Wenn Verfahren zur Personalauswahl eingesetzt werden, so verspricht man sich davon Hinweise, die es tatsächlich erlauben, die geeignetsten Kandidatinnen bzw. die geeignetsten Kandidaten aus den Bewerberinnen und Bewerbern heraus zu suchen. Die Verfahren sollen also im weitesten Sinne die „Eignung“ feststellen. Ob ein Verfahren tatsächlich das misst, was es zu messen vorgibt, hier die „Eignung“, wird als *Validität* des Verfahrens bezeichnet. Zur Feststellung der Validität wird in der Regel eine ↗ Korrelation zwischen den Ergebnissen des eingesetzten Verfahrens und passender Außenkriterien (z.B. Leistungsbeurteilung durch Vorgesetzte) berechnet. Damit ist die Validität quantifizierbar mit Werten zwischen Null und Eins, wobei hohe Werte einer hohen Validität entsprechen. Da es bei der Personalauswahl darum geht, die Eignung zu prognostizieren und als passende Außenkriterien Merkmale in Frage kommen, die in der Zukunft liegen, spricht man von einer prognostischen Validität, also von der Fähigkeit des eingesetzten Verfahrens, Vorhersagen über die Verwendbarkeit einer Bewerberin eines Bewerbers zu erstellen. Wie hoch die Validität im Idealfall sein soll, hängt vom Einsatzziel (z.B. von der Anzahl der wahrscheinlich ohnehin geeigneten Bewerberinnen und Bewerbern: sind wahrscheinlich ohnehin alle für die Stelle geeignet, kann die Auswahl einfach gehalten werden) und vom Aufwand (Kosten vs. Nutzen) ab. Eine hohe Validität wird Verfahren mit einem Wert über 0,3 zugesprochen. Hierzu gehört z.B. das Assessment Center, wohingegen Bewerbungsunterlagen, Schulnoten und graphologische Gutachten darunter liegen.

Varianzanalyse. (heißt auch ANOVA) In der Regel sind Signifikanztests in der Lage nur zwei Kennwerte (z.B. Mittelwerte) miteinander zu vergleichen. Einige Fragestellungen machen daher mehrere Vergleiche zwischen jeweils zwei Messwerten nötig, um die Frage insgesamt beantworten zu können. Beantworten drei Personengruppen einen Fragebogen (Gruppe A, B, C), so kommt man auf insgesamt drei paarweise Vergleiche (A mit B; A mit C und B mit C). Obwohl es hier möglich ist, jede Kombination der Gruppen einzeln zu vergleichen und eine Alpha-Fehler-Adjustierung vorzunehmen (↗ Alpha-Fehler-Adjustierung), ist eine Varianzanalyse eleganter und weniger aufwändig zu rechnen. Die Varianzanalyse löst das Problem durch einen Trick: Es werden im Wesentlichen zwei Varianzen (↗ Standardabweichung, Streuung, Varianz) ermittelt und diese mit einem F-Test verglichen. Es werden also auch hier nur zwei Kennwerte (hier Varianzen) durch den Test verglichen. Die eine Varianz ist die innerhalb der Gruppen, die andere ist die zwischen den Gruppen. Sind die Unterschiede (ermittelt durch die Varianz) zwischen den Gruppen größer als die Unterschiede innerhalb der Gruppen, so unterscheiden sich die Gruppen. Allerdings ist dann noch nicht bekannt, welche der Gruppen sich voneinander unterscheiden. Um dies herauszufinden werden anschließend doch wieder paarweise Vergleiche durchgeführt. Für eine Varianzanalyse werden klar abgegrenzte Untersuchungsgruppen (↗ Kategorien) benötigt. Die Regressionsanalyse ist eine flexiblere Alternative, wenn solche Kategorien nicht vorliegen.

Zentraler Grenzwertsatz. Ist eine Untersuchungsstichprobe groß, so ergibt sich unabhängig von der Verteilung der Rohdaten für den Mittelwert eine Normaverteilung. Diesen Zusammenhang kann man sich wie folgt vorstellen: Es wird aus einer größeren Stichprobe eine begrenzte Zufallsauswahl getroffen und für diese Zufallsauswahl ein Mittelwert berechnet. Dies wird mehrfach wiederholt. Jeder berechnete Mittelwert beruht dann nur auf einer Zufallsauswahl und stimmt damit mit dem echten Mittewert nur mehr oder weniger gut überein. Es zeigt sich, dass die Mittelwerte der Zufallsauswahlen um den echten Mittelwert normalverteilt streuen und zwar unabhängig von der Verteilung der eigentlichen Rohwerte. Testverfahren wie der ↗ T-Test oder die ↗ Varianzanalyse benötigen solche normalverteilten Mittelwerte. Diese sind nach dem zentralen Grenzwertsatz für große Stichproben immer gegeben. Was groß ist und was nicht hängt vom jeweiligen Lehrbuch ab. Einige sagen, dass 25 Personen pro Untersuchungs-Gruppe genügen, andere fordern 30 und ganz strenge sogar 50 Messwerte. Da für kleine Stichproben der zentrale Grenzwertsatz nicht gilt, kann eine Normalverteilung der Mittelwerte nur dann erwartet werden, wenn auch die Rohwerte normalverteilt sind. Dies muss für kleine Stichproben geprüft werden. Bei großen Stichproben ist eine solche Prüfung irreführend und sollte unterbleiben.

Darstellung und Abkürzungen

Allgemein.

AM *oder* M *oder* $\bar{x}$	*Mittelwert (arithmetisches Mittel; Mean)*
SD *oder* s *oder* Std.	*Standardabweichung (Standard Deviation; Streuung)*
df *oder* FG	*Freiheitsgrade (degree of freedom)*
N	*Größe der erfassten Grundgesamtheit.*
n	*Größe einer Stichprobe oder Gruppe aus der Grundgesamtheit.*
p	*Wahrscheinlichkeit (kann Werte zwischen 0 und 1 annehmen. 0,6 bedeutet also eine Wahrscheinlichkeit von 60%).*
p-2-seitig	*Wahrscheinlichkeit dafür, dass etwas nicht signifikant ist (2-seitig getestet).*
p-1-seitig	*Wahrscheinlichkeit dafür, dass etwas nicht signifikant ist (1-seitig getestet).*
*	*Der Unterschied ist signifikant bei einem Alphafehler von 5% (p <= 0,05)*
**	*Der Unterschied ist hoch signifikant bei einem Alphafehler von 1% (p <= 0,01)*

Korrelationen.

r	*Korrelation.*
p-2-seitig	*Wahrscheinlichkeit dafür, dass eine Korrelation Null ist.*
*	*Die Korrelation ist mit einem Alphafehler (einer Wahrscheinlichkeit) von 5% (p <= 0,05) Null. Die Korrelation ist signifikant.*
**	*Die Korrelation ist mit einem Alphafehler (einer Wahrscheinlichkeit) von 1% (p <= 0,01) Null. Die Korrelation ist hoch signifikant.*

Im Text: *„...Es besteht also ein hoch signifikanter Zusammenhang zwischen dem Verhalten der beobachteten Frauen und dem Verhalten der beobachteten Männer (r = 0,52; p-2-seitig = 0,001)....“*

In Tabellen:

Tabelle 1: Interkorrelationsmatrix der Skalen

N = 506	**Skala 1**	**Skala 2**	**Skala 3**	**Skala 4**	**Skala 5**
Skala 2	0,758 **				
Skala 3	0,924 **	0,723 **			
Skala 4	0,815 **	0,589 **	0,292 *		
Skala 5	0,810 **	0,491 **	0,587 **	0,517 **	
Skala 6	0,849 **	0,599 **	0,062	0,706 **	0,562 **

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.
Skala 1: Neurotizismus
Skala 2: Emotionale Stabilität
Skala 3: Extraversion
Skala 4: Selbstdarstellung
Skala 5: Führungsmotivation

T-Test.

t	*Prüfgröße für den T-Test.*
df	*Freiheitsgrade (degree of freedom).*
p-2-seitig	*Wahrscheinlichkeit dafür, dass zwei Mittelwerte sich nicht signifikant unterscheiden (2-seitig getestet).*
p-1-seitig	*Wahrscheinlichkeit dafür, dass zwei Mittelwerte sich nicht signifikant unterscheiden (1-seitig getestet).*
*	*Der Unterschied ist signifikant bei einem Alphafehler von 5% (p <= 0,05)*
**	*Der Unterschied ist hoch signifikant bei einem Alphafehler von 1% (p <= 0,01)*

Im Text: *„...Es besteht also ein hoch signifikanter Unterschied zwischen dem Verhalten der beobachteten Frauen und dem Verhalten der beobachteten Männer (t = 3,52; df = 255; p-2-seitig = 0,003)....“*

In Tabellen:

Tabelle 2: Ergebnisse der Befragung: Männer vs. Frauen[1]

	Männer			Frauen					
	AM	**SD**	**n**	**AM**	**SD**	**n**	**t**	**df**	**p**
Skala 1	5,25	1,32	500	6,01	1,12	420	2,57	918	0,004**
Skala 2	4,98	1,04	499	5,98	1,20	420	2,62	917	0,003**

** Die Unterschiede sind auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Unterschiede sind auf dem Niveau von 0,05 (2-seitig) signifikant.
Skala 1: Neurotizismus
Skala 2: Emotionale Stabilität

Tabelle 2 (alternative): Ergebnisse der Befragung: Männer vs. Frauen[2]

	Männer		Frauen				
	AM (SD)	**n**	**AM (SD)**	**n**	**t**	**df**	**p**
Skala 1	5,25 (1,32)	500	6,01 (1,12)	420	2,57	918	0,004**
Skala 2	4,98 (1,04)	499	5,98 (1,20)	420	2,62	917	0,003**

** Die Unterschiede sind auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Unterschiede sind auf dem Niveau von 0,05 (2-seitig) signifikant.
Skala 1: Neurotizismus
Skala 2: Emotionale Stabilität

[1] Alle Zahlenwerte (inklusive der Ergebnisse des Tests) sind frei erfunden
[2] Alle Zahlenwerte (inklusive der Ergebnisse des Tests) sind frei erfunden

Varianzanalyse.

F	*Prüfgröße für den F-Test, der in der Varianzanalyse benutzt wird*
df	*Freiheitsgrade (degree of freedom).*
p	*Wahrscheinlichkeit dafür, dass sich die Gruppen nicht signifikant unterscheiden*
*	*Der Unterschied zwischen den Gruppen ist signifikant bei einem Alphafehler von 5% (p <= 0,05)*
**	*Der Unterschied zwischen den Gruppen ist hoch signifikant bei einem Alphafehler von 1% (p <= 0,01)*

Im Text: *„...Es besteht also ein hoch signifikanter Unterschied zwischen den beobachteten Gruppen (F = 3,52; p = 0,003)....“*

In Tabellen:

Tabelle 3:
Ergebnisse der Beobachtung[3]

***Skala* 1**	**AM**	**SD**	**N**	**F**	**p**
Männer	5,25	1,32	500	3,25	0,004**
Frauen	4,98	1,04	499		
Kinder	3,12	0,98	356		
***Skala* 2**					
Männer	4,25	1,32	500	2,56	0,003**
Frauen	4,98	1,04	499		
Kinder	3,12	0,98	356		

** Die Unterschiede sind auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Unterschiede sind auf dem Niveau von 0,05 (2-seitig) signifikant.
Skala 1: Neurotizismus
Skala 2: Emotionale Stabilität

Tabelle 4: Signifikante Gruppenunterschiede nach Scheffé-Test[4]

***Skala* 1**	Frauen	Kinder
Männer	*	**
Frauen		**
***Skala* 2**		
Männer	*	
Frauen		**

** Die Unterschiede sind auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Unterschiede sind auf dem Niveau von 0,05 (2-seitig) signifikant.
Skala 1: Neurotizismus
Skala 2: Emotionale Stabilität

Bei einer Varianzanalyse bieten sich zusätzlich auch Grafiken an.

[3] Alle Zahlenwerte (inklusive der Ergebnisse des Tests) sind frei erfunden
[4] Alle Zahlenwerte (inklusive der Ergebnisse des Tests) sind frei erfunden

Literatur

Cohen, J. (1992) A Power Primer. *Psychological Bulletin*, 112 (1), 155-159