Mittelwertvergleiche mittels t-Test

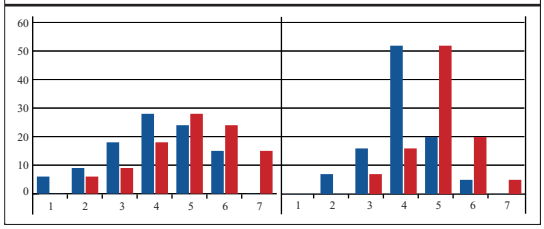
Der t-Test für unverbundene beziehungsweise unabhängige Stichproben zählt zu den am häufigsten genutzten statistischen Tests. Er überprüft, ob sich die Mittelwerte metrischer Merkmale in zwei Test- oder Teilgruppen signifikant voneinander unterscheiden.

Einführungsbeispiel. Auf einer 7-stufigen Rating-Skala wurde die Kaufabsicht für ein neues Produkt erhoben. Die durchschnittliche Kaufabsicht beträgt in der (Teil-)Stichprobe der weiblichen Befragten 5, in der der männlichen Befragten 4. Ist der Unterschied signifikant? Auskunft darüber gibt der p-Wert als Ergebnis eines geeigneten statistischen Tests. Der t-Test für unverbundene Stichproben ist geeignet, da die Gruppen der Frauen und Männer sich nicht überschneiden. Er geht von der Hypothese aus, dass die Mittelwerte in den Grundgesamtheiten gleich sind. Der p-Wert gibt die Wahrscheinlichkeit an, mit dem Verwerfen dieser Hypothese einen Fehler zu begehen. Ist der p-Wert kleiner als ein vorgegebenes Signifikanzniveau α (zumeist ist $\alpha = 0{,}05$), so wird die Hypothese gleicher Mittelwerte abgelehnt, weil die Wahrscheinlichkeit sehr gering ist, mit dieser Entscheidung falsch zu liegen. Der Unterschied wäre dann signifikant.

Konstruktion der Teststatistik. Jedem statistischen Test liegt eine Teststatistik zugrunde, deren Wert auf Basis der Stichprobe berechnet wird. Dieser wiederum determiniert den p-Wert. Für den Vergleich der Mittelwerte liegt es nahe, dass in die Teststatistik die Differenz der Mittelwerte eingeht. Trotz gleicher Differenzen können sich jedoch die p-Werte und somit auch die Beurteilungen im Hinblick auf die Signifikanz unterscheiden.

Die Abbildung zeigt die Häufigkeitsverteilungen der Antworten auf die Kaufabsicht in zwei Untersuchungen. In beiden ist der Mittelwert für die blaue Gruppe 4 und für die rote Gruppe 5. Allerdings ist die Streuung der Antworten der Untersuchung im linken Teil der Abbildung größer als im rechten. Die blaue und rote Gruppe scheinen sich im Fall mit geringerer Streuung deutlicher zu unterscheiden, da sich die beiden Verteilungen weniger überlappen. Daher wird die Differenz der Mittelwerte $\bar{x}_1 - \bar{x}_2$ in Relation zur Standardabweichung s des Merkmals beurteilt.

Häufigkeitsverteilungen der Antworten von jeweils zwei Teilgruppen



Bei im Vergleich zur Grundgesamtheit großen Stichproben ist die Gefahr kleiner, dass die Mittelwerte der Stichproben sehr von den Mittelwerten der Grundgesamtheiten abweichen, als bei kleineren Stichproben. Um dies zu berücksichtigen, wird der Quotient aus Mittelwertdifferenz und Standardabweichung mit einem Faktor multipliziert, der die Teilstichprobenumfänge n_1 und n_2 beinhaltet:

$$t = \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \cdot \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Die gemeinsame Standardabweichung s ergibt sich aus den mit n_1 und n_2 gewichteten Standardabweichungen innerhalb der Teilstichproben. Die Teststatistik t folgt unter der Annahme der Gültigkeit der Hypothese gleicher Mittelwerte in den Grundgesamtheiten dann einer t- beziehungsweise Student-Verteilung, die der Normalverteilung sehr ähnlich ist und dem Test seinen Namen gibt.

Eine vorliegende Stichprobe ist immer nur eine von sehr vielen desselben Umfangs, die sich bei einer Zufallsauswahl ebenso hätten ergeben können. Der p-Wert ist die Wahrscheinlichkeit, eine Stichprobe zu ziehen, für die der Wert der Teststatistik vom Betrag größer ist als der der tatsächlich gezogenen Stichprobe. Je höher der berechnete t-Wert ist, desto geringer ist die Wahrscheinlichkeit, dass der einer anderen Stichprobe größer ist. Das heißt ein Unterschied zwischen zwei Mittelwerten ist umso eher signifikant,

- je größer die Differenz der Mittelwerte ist,
- je kleiner die Standardabweichung ist und
- je größer die Umfänge der Teilstichproben sind.

Voraussetzungen. Die Größe der beiden Gruppen sollte jeweils mindestens 30 betragen, wenn nicht von einer Normalverteilung des Merkmals in den Grundgesamtheiten ausgegangen werden kann. Anderenfalls ist ein Test aus der Gruppe der nicht-parametrischen (verteilungsfreien) Verfahren zu nutzen.

Sind die Stichprobenumfänge unterschiedlich hoch, beeinträchtigt dies kaum die Testergebnisse, solange die Varianzen in den Grundgesamtheiten als gleich gelten können. Sind Stichprobenumfänge und Varianzen in den beiden Gruppen verschieden, ist eine Variante des beschriebenen Tests, der Welch-Test, einzusetzen.

In Ausgabe 3/2016: Effektstärke

Die Autoren

Johannes Lüken, Diplom-Psychologe, ist Leiter des Bereichs Data Sciences bei IfaD.

jlueken@ifad.de

Prof. Dr. Heiko Schimmelpfennig, Diplom-Kaufmann, ist Projektleiter für Data Sciences bei IfaD.

hschimmelpfennig@ifad.de

Literatur

Bortz, J.; Schuster, C. (2010): ***Tests zur Überprüfung von Unterschiedshypothesen.*** In: Statistik für Human- und Sozialwissenschaftler, Berlin, S. 117–136

Lüken, J.; Schimmelpfennig, H. (2012): ***Einführung in Signifikanztests.*** In: planung&analyse, Nr. 5/2012, S. 24