Ein genauerer Blick auf die Empirie in Frau Raabs Diplomarbeit

verfasst von Mag.Mag.Dr.sc.ETH Sabrina Dorn, MSc ETH am 11.1.2022

Nicht nur zahlreiche Plagiate in Susanne Raabs Diplomarbeit erwecken den Eindruck von Quatsch, sondern auch die in Abschnitt 6 dargestellten empirischen Ergebnisse machen es einem wahrlich schwer sich der Schlussfolgerung zu entziehen, dass Frau Raab und vor allem aber auch der mittlerweile verstorbene Betreuer ihrer Diplomarbeit - zumindest zum Zeitpunkt als die Arbeit verfasst wurde - nicht allzu viel Ahnung von Statistik hatten. Diese Kritik bezieht sich einerseits auf die korrekte Anwendung von statistischen Methoden und einen teilweisen Mangel an intersubjektiver Nachvollziehbarkeit, wie auch andererseits auf fehlerbehaftete Berechnungen und Interpretationen. Wie man unweigerlich zu diesen Schlussfolgerungen kommen muss, sei im Folgenden eingehender erläutert.

1. Methoden

In Abschnitt 6.1 der Diplomarbeit werden die verwendeten Methoden zur statistischen Auswertung dargestellt. Ganz grundsätzlich macht es für die im Rahmen der Arbeit erhobenen Daten durchaus Sinn, Mittelwerte aus zwei Stichproben mit einem Zwei-Stichproben-t-Test und Mittelwerte über mehrere Gruppen hinweg mit einem F-Test (Verallgemeinerung des t-Tests; ANOVA) zu vergleichen, vorausgesetzt die Annahmen für diese Tests sind seitens der Daten auch tatsächlich erfüllt. Werden nämlich statistische Verfahren angewendet, dann entspricht es einer guten wissenschaftlichen Praxis auch zu überprüfen ob die Annahmen für deren Validität seitens der zu analysierenden Daten erfüllt sind und dies im Rahmen einer entsprechenden wissenschaftlichen Arbeit auch zu erläutern bzw. gegenüber der Leserin zu verifizieren (intersubjektive Nachvollziehbarkeit). Sind wesentliche Annahmen nicht erfüllt, so ist eine Verzerrung der gewonnen Schätzer eine wahrscheinliche Folge. Sowohl der Zwei-Stichproben-t-Test als auch die ANOVA setzen voraus, dass die zugrunde liegenden Daten von einander unabhängigen, jeweils unabhängig normal-verteilten Zufallsstichproben mit gleichen Varianzen entstammen. Dass eine Zufallsvariable innerhalb einer Stichprobe unabhängig verteilt ist, kann im Rahmen einer Fragebogenerhebung durch ein entsprechend sauberes Design bei der Erhebung sicher gestellt werden, die anderen beiden Annahmen können durch entsprechende statistische Tests oder eine graphische Analyse - und im besten Fall mittels beidem - überprüft werden. Details dazu können etwa in [1] in Abschnitt 2.4 zum Zwei-Stichproben-t-Test und in Abschnitt 3.4 zur einfaktoriellen ANOVA bzw. in [2] nachgelesen werden. Einzig auf Normalverteilung der zu Grunde liegenden Größen zu testen - wie in Anhang III der Diplomarbeit - ist in diesem Zusammenhang jedenfalls unzureichend für eine solide statistische Argumentation.

Auf Seite 52 unter Punkt 6.1.2 beschreibt die Autorin das Verfahren des Kolmogorov-Smirnov-Tests. Dieser nicht-parametrische Test wird in der Arbeit angewendet um zu

testen, ob eine empirische Größe aller Wahrscheinlichkeit nach normalverteilt ist. Bei der Beschreibung wurde der Satz

„Grundlage des Tests ist die Berechnung der maximalen Differenz zwischen den kumulierten Häufigkeiten beider Stichproben."

nicht nur abgeschrieben, sondern eben auch falsch abgeschrieben. In der Diplomarbeit wurde der Ein-Stichproben-Kolmogorov-Smirnov-Test angewendet um eine empirische Verteilung mit einer Normalverteilung zu vergleichen. Es gibt allerdings auch einen Zwei-Stichproben-Kolmogorov-Smirnov-Test mit welchem zwei empirische Verteilungen aus unterschiedlichen Stichproben dahingehend miteinander verglichen werden, ob diese aller Wahrscheinlichkeit nach identisch sind und genau darauf bezieht sich der abgeschriebene Satz.

Auch bei der Beschreibung des Mann-Whitney-U-Tests auf Seite 53 unter Punkt 6.1.7 hat die Autorin nicht nur abgeschrieben, sondern auch unvollständig. So steht dort geschrieben

„Dieses Verfahren wird angewendet, um zu überprüfen, ob zwei unabhängige Stichproben aus einer gemeinsamen Grundgesamtheit entstammen oder nicht. Ein signifikanter Wert zeigt an, dass sich die zwei unabhängigen Stichproben voneinander unterscheiden. Sie entstammen somit nicht einer Grundgesamtheit."

Mit dem U-Test können im Rahmen eines verteilungsfreien Verfahrens zwei Gruppen dahingehend untersucht werden, ob sich ihre sogenannten zentralen Tendenzen (Mittelwert bzw. Median) unterscheiden, vorausgesetzt die Verteilungen sind gleich bis auf Unterschiede im zentralen Lagemaß (vgl. [3]). Sehen die Verteilungen in den verglichenen Gruppen aber sehr unterschiedlich aus, so ist die Interpretation unter der Alternativhypothese, dass die beiden Verteilungen nicht der gleichen Grundgesamtheit entstammen, zulässig (vgl. [4]). Welcher Fall erfüllt ist, sollte aber vorab etwa mit einer graphischen Gegenüberstellung der empirischen Dichten bzw. Verteilungsfunktionen überprüft werden (vgl. [3] und [4]). Ein Mann-Whitney-U-Test wurde in der Diplomarbeit von Frau Raab für die Tabellen 25 und 33 durchgeführt und für Unterschiede im Lagemaß interpretiert (siehe Seiten 69 und 76) ohne vorab zu überprüfen, ob sich die Verteilungen in den beiden Gruppen tatsächlich nur hinsichtlich Lage, nicht jedoch hinsichtlich From unterscheiden. Warum in diesen beiden Fällen ein U- anstatt eines t-Tests verwendet wird, wird von Frau Raab grundsätzlich korrekt damit begründet, dass in diesen Fällen die Nullhypothese der Normalverteilung des Kolmogorov-Smirnov-Tests jeweils auf einem Signifikanzniveau von 5% verworfen wurde. Dazu wurden in Anhang III ihrer Diplomarbeit insgesamt 108 solcher Tests durchgeführt. Das Signifikanzniveau (oder Typ-I-Fehler) eines statistischen Tests gibt die Wahrscheinlichkeit an, eine wahre Nullhypothese zu verwerfen. Bei 108 durchgeführten Kolmogorov-Smirnov-Tests würde man sich folglich erwarten, dass 5.4 mal eine wahre Nullhypothese verworfen wird. Tatsächlich wurde in Frau Raabs

Arbeit die Nullhypothese in 3 von 108 Fällen auf einem Signifikanzniveau von 5% verworfen. Man hätte hier aufgrund des wiederholten Testens ein geringeres Signifikanzniveau zu Grunde legen sollen um die Nullhypothese der Normalverteilung zu verwerfen. Weiters zeigt dies ganz klar, wie unabkömmlich graphische Analysen in Bezug auf statistische Modellannahmen sind, was zumindest der Betreuer einer empirischen Diplomarbeit hätte wissen müssen.

Weiters muss es bei der Interpretation insbesondere der Ergebnisse der durchgeführten Zwei-Stichproben-t-Tests sowohl bei Frau Raab als auch bei ihrem Betreuer gröbere Verständnisprobleme gegeben haben: Auf Seite 47 unter Punkt 5 schreibt die Autorin, dass die zweiseitige der einseitigen Hypothesenformulierung und -prüfung vorgezogen wurde, jedoch steht dazu in Widerspruch, dass die Ergebnisse der verwendeten t-Tests durchgehend wie bei einer einseitigen Alternativhypothese interpretiert wurden (siehe Seiten 62, 66, 71, 72, 76, 80 und 82). Auch die durchgeführten deskriptiven Mittelwertvergleiche sind statistisch ohne jeglichen Gehalt (vgl. Abbildungen 7, 8, 9, 10 usw.). Sollte es tatsächlich die Intention der Autorin gewesen sein hier einseitige Alternativhypothesen zu berücksichtigen, wurden die p-Werte für die Zwei-Stichproben-t-Tests schlichtweg falsch berechnet.

2. Replikation einzelner Ergebnisse

Für die Zwei-Stichproben-t-Tests und die ANOVA erlauben es die in Anhang III der Diplomarbeit (siehe Seiten 104 ff.) gegebenen Mittelwerte und Standardabweichungen sowie Stichprobengrößen die in Abschnitt 6 präsentierten empirischen Ergebnisse auch ohne Vorliegen der Rohdaten der Umfrage einer Replikationsprüfung zu unterziehen. Ausgenommen sind die Ergebnisse aus Tabellen 26 und 28, da hier auch keine Kolmogorov-Smirnov-Tests gerechnet wurden und die entsprechenden Parameter für die Berechnung der Zwei-Stichproben-t-Tests und der ANOVA nicht gegeben sind. Zusätzlich wurde im Rahmen dieser Replikationsprüfung auch die Annahme gleicher Varianzen mit einem F- bzw. Bartlett-Test überprüft. Zum F-Test auf gleiche Varianzen siehe z.B. [5] und zum Bartlett-Test siehe z.B. Seite 79 in [1]. Die im Detail durchgeführten Berechnungen sind hier im Anhang aufgeführt. Wie im Folgenden in 2.1 und 2.2 erläutert, erwiesen sich 7 von 13 replizierten Berechnungen als fehlerbehaftet.

2.1 Zwei-Stichproben-t-Tests

Tabelle 1 fasst die Ergebnisse der Replikationsprüfung der Zwei-Stichproben-t-Tests zusammen. In 3 von 8 Fällen muss dabei die Nullhypothese gleicher Varianzen auf Basis eines F-Tests zugunsten einer zweiseitigen Alternativhypothese verworfen werden (betrifft die Tabellen 18, 21 und die Variable „Altruismus“ aus Tabelle 27), sodass eine grundlegende Annahme für die Validität des Zwei-Stichproben-t-Test verletzt ist. Folglich

hätte in der Diplomarbeit in diesen Fällen gar kein Zwei-Stichproben-t-Test zur Anwendung kommen dürfen, sondern es hätte etwa ein Welch-Test mit einer Satterthwaite-Approximation für die Freiheitsgrade - wie hier angewendet - herangezogen werden können. Weiters stimmen die in den Tabellen 17 und 18 der Diplomarbeit angegebenen Mittelwerte nicht mit den in Anhang III gegeben Werten für dieseleben Parameter überein, sodass sich auch hier Fehler in die in der Diplomarbeit präsentierten statistischen Ergebnisse eingeschlichen haben. Für die Variable „Egoismus“ im unteren Teil von Tabelle 27 wurde der p-Wert für die Teststatistik ausnahmsweise der dort verwendeten Interpretation entsprechend tatsächlich für einen einseitigen Test berechnet, jedoch ist unklar ob dies so beabsichtigt war. Schlussfolgernd sind damit 5 von 8 replizierten Zwei-Stichproben-t-Tests fehlerbehaftet.

Tabelle 1: Ergebnisse der Replikationsprüfung der t-Tests und Gegenüberstellung

	Werte laut Diplomarbeit		Überprüfung der Annahme gleicher Varianzen		Replizierte Werte		
Tabelle	Teststatistik	p-Wert	F-Test	p-Wert	Teststatistik	p-Wert	Test
17	-1.66	0.101	0.757	0.353	-0.661	0.510	t
18	-0.661	0.510	1.864	0.048	-1.657	0.098	Welch
21	5.637	0.000	0.515	0.028	5.637	0.000	Welch
27: Altruismus	2,620	0.011	0.494	0.020	2.620	0.010	Welch
27: Egoismus	0,060	0.4625	1.108	0.748	0.060	0.952	t
32	-2,948	0.004	0.953	0.865	-2.948	0.004	t
38	-2,76	0.007	1.669	0.103	-2.764	0.007	t
41	-2,878	0.004	1.629	0.120	-2.978	0.004	t

2.2 ANOVA

Tabelle 2 fasst die Ergebnisse der Replikationsprüfung der in der Arbeit verwendeten ANOVAs (soweit aus den Daten in Anhang III der Arbeit replizierbar) zusammen. Es wurde jeweils die Annahme gleicher Varianzen mit Hilfe eines Bartlett-Tests überprüft mit dem Ergebnis, dass die Nullhypothese auf herkömmlichen Signifikanzniveaus nicht verworfen werden konnte. Damit ist diese Annahme für die Anwendbarkeit des F-Tests aller Wahrscheinlichkeit nach erfüllt, obwohl zu bedenken bleibt, dass die Teilstichproben jeweils relativ kleine Beobachtungszahlen aufweisen und zumindest seitens des Betreuers dieser Arbeit auf die Möglichkeit einer graphischen Analyse dieser Annahme hingewiesen hätte werden müssen. Es sei auch noch erwähnt, dass die Software SPSS - welche in der Diplomarbeit zur Auswertung der Daten herangezogen wurde - im Output für die einfaktorielle ANOVA automatisch eine Levene-Statistik zwecks Überprüfung der Annahme

der Varianzhomogenität mitausgibt (siehe [6]). Folglich hätte man zwecks intersubjektiver Nachvollziehbarkeit diese nur in der Arbeit wiedergeben und interpretieren müssen. Wie aus Tabelle 2 ersichtlich, kam es im Rahmen der Replikationsprüfung für 2 von 5 der ANOVAs zu abweichenden Ergebnissen.

Tabelle 2: Ergebnisse der Replikationsprüfung der ANOVA und Gegenüberstellung

	Quadratsummen		Mittel der Quadrate			
Tabelle	Zwischen	Innerhalb	Zwischen	Innerhalb	F	p-Wert
19	59.399	271.245	19.800	3.154	6.278	0.001
repliziert	29.541	361.804	9.847	4.207	2.341	0.079
22	1938.539	1907.283	646.180	22.178	29.136	0,000
repliziert	215.387	211.921	71.796	2.464	29.136	0.000
34	249.937	1484.963	83.312	17.267	4.825	0.004
repliziert	249.937	1484.962	83.312	17.267	4.825	0.004
34	422.198	1715.625	140.733	19.949	7.055	0,000
repliziert	422.200	1715.626	140.733	19.949	7.055	0.000
39	206,788	1447,668	68,929	16,833	4,095	0,009
repliziert	206.788	1447.668	68.930	16.833	4.095	0.009

3. Raten statt testen macht man in der Wissenschaft nicht!

Als ein Ziel der empirischen Untersuchung wird in der Diplomarbeit genannt, zu untersuchen, ob es signifikante Unterschiede hinsichtlich Lebensbedeutung und Einstellungsstrukturen der erhobenen Stichprobe zur Allgemeinbevölkerung gibt. Dazu wurden die erhobenen Daten mit einer Normierungsstichprobe (LEBE) verglichen. Eine vertiefende Untersuchung dieser Frage unterblieb allerdings, sondern sie wurde nur sehr rudimentär an zwei Stellen der Arbeit (siehe Tabellen 30 und 31 sowie die Tabellen 36 und 37) behandelt. Statistische Test wurde hierzu allerdings verwunderlicherweise nicht durchgeführt. Im Folgenden werden daher - soweit ohne die Rohdaten der Umfrage möglich - die diesbezüglich in der Arbeit getroffenen qualitativen Aussagen mittels statistischer Tests überprüft:

Da die geschlechterspezifischen Standardabweichungen zu Tabelle 30 in der Arbeit nicht angegeben sind, können die Schlussfolgerungen zu den Tabellen 30 und 31 keiner vollständigen Replikationsprüfung unterzogen werden. Für die gesamte Stichprobe (Männer und Frauen) ist es allerdings möglich die Standardabweichung aus Anhang III zu replizieren. Ein F-Test auf gleiche Varianzen für die Variable „Soziales Engagement“ zeigt, dass hier die Nullhypothese gleicher Varianzen zugunsten einer zweiseitigen

Alternativhypothese verworfen werden muss (p-Wert 0.002). Die Anwendung eines Welch-Tests der Nullhypothese, dass es keine Unterschiede zur Allgemeinbevölkerung gibt, zeigt, dass diese nicht verworfen werden kann (zweiseitiger p-Wert 0.578). Auch für die Variable „Generativität“ muss auf Basis eines F-Tests der Nullhypothese der Varianzhomogenität zugunsten einer zweiseitigen Alternativhypothese verworfen werden (p-Wert 0.002). Der im Anschluss durchgeführte Welch-Test der Nullhypothese, dass es hier keine Unterschiede zur Allgemeinbevölkerung gibt, kann nicht verworfen werden (zweiseitiger p-Wert 0.739). Obwohl die in der Diplomarbeit getroffenen Schlussfolgerungen dort nicht durch eine entsprechende statistische Untersuchung belegt sind, können diese damit inhaltlich bestätigt werden.

Zu Tabellen 36 und 37 folgert Frau Raab, dass Mitglieder der ELSA überdurchschnittliche Werte auf den Skalen „Leistung“ und „Macht“ im Vergleich zur Normierungsstichprobe aufweisen. Diese Aussage kann allerdings im Rahmen einer überprüfenden statistischen Testung nicht in der Form belegt werden.

Unter Verwendung der Daten aus Anhang III wurde dazu ein Zwei-Stichproben-t-Test für die Variable „Leistung“ berechnet (die Nullhypothese gleicher Varianzen der ELSA-Stichprobe und der Normierungsstichprobe kann mittels eines vorher durchgeführten F-Tests nicht verworfen werden) mit dem Ergebnis, dass die Nullhypothese, dass die Variable „Leistung“ unter Freiwilligen, die sich bei ELSA engagieren, und der Normierungsstichprobe gleich ist, nicht verworfen werden kann (zweiseitiger p-Wert 0.120). Auch für Freiwillige, die sich bei der Fachschaft engagieren, kann die Nullhypothese der Varianzhomogenität nicht verworfen werden. Der im Anschluss durchgeführte t-Test zeigt, dass die Nullhypothese keiner Unterschiede zur Allgemeinbevölkerung bei Zugrundelegung eines Signifkanzniveaus von 5% knapp nicht verworfen werden kann (zweiseitiger p-Wert 0.051). Schlussfolgernd kann damit die in der Diplomarbeit getroffene qualitative Aussage betreffend der Variable „Leistung“ nicht statistisch belegt werden.

Für die Variable „Macht“ wurden entsprechende Berechnungen durchgeführt. Die Annahme gleicher Varianzen - jeweils im Vergleich zur Normierungsstichprobe - kann weder für Freiwillige bei ELSA noch bei der Fachschaft verworfen werden, sodass in beiden Fällen ein Zwei-Stichproben-t-Test herangezogen werden kann. In beiden Fällen (Freiwillige bei ELSA vs. Normierungsstichprobe und Freiwillige bei der Fachschaft vs. Normierungsstichprobe) muss die Nullhypothese zu Gunsten der zweiseitigen Alternativhypothese verworfen werden (zweiseitige p-Werte in beiden Fällen < 0.01). Schlussfolgernd unterscheiden sich damit beide Stichproben hinsichtlich der Variable „Macht“ signifikant von der Allgemeinbevölkerung und nicht nur für Freiwillige bei ELSA, wie in der Diplomarbeit behauptet.

Literaturverzeichnis

[1] Montgomery, Douglas. C., Design and Analysis of Experiments, 7. Auflage, Wiley and Sons (2009)
[2] Kalisch, Markus, Analysis of Variance, Vorlesungsunterlagen der ETH Zürich (2014), https://stat.ethz.ch/education/semesters/as2014/statistik2/v5.pdf, abgerufen am 11.1.2022

[3] Müller Patric, Hypothesentests für zwei Stichproben, Vorlesungsunterlagen der ETH Zürich (2019), https://ethz.ch/content/dam/ethz/special-interest/math/statistics/sfs/Education/Advanced%20Studies%20in%20Applied%20Statistics/course-material-1921/W%27keit%20und%20Statistik/slides09.pdf, abgerufen am 10.1.2022

[4] The ultimate IBM® SPSS® Statistics guides (2018), https://statistics.laerd.com/statistical-guides/mann-whitney-u-test-assumptions.php, abgerufen am 11.1.2022

[5] Methodenberatung der Universität Zürich (2020), https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/unterschiede/varianzen/ftest.html, abgerufen am 10.1.2022

[6] Janssen, Jürgen und Laatz Wilfried, Statistische Datenanalyse mit SPSS - Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests, 8. Auflage, Springer Gabler (2013)

Anhang (siehe separates Dokument)