

Statistiksupport bei Diplomarbeit

(Dieses und weitere Papiere finden Sie auf <http://sozio.logik.ch> unter „Statistiksupport PH Wallis“)

Die PH Wallis bietet den Studierenden, welche bei der Diplomarbeit statistische Methoden anwenden, einen Support an. Der Support übernimmt

- auf der Basis von Daten deren Auswertung (beschreibende und schliessende Statistik) und
- berät die Studierenden bei statistischen Fragen.
- Er übernimmt auf Wunsch die Aufgabe, Fragebogen aufs Internet zu stellen.
- Er ist bereit, vor der Abgabe der Diplomarbeit, Statistikpassagen der Arbeit auf Korrektheit zu überprüfen.

Es ist aber nicht die Aufgabe des Statistiksupports, Hypothesen aufzustellen zu helfen oder sonstige inhaltliche Fragen zu diskutieren (Aufgabe des Betreuers der Diplomarbeit; Ausnahme: der Betreuer verweist die Diplomandin für solche Fragen an den Statistiksupport).

Das vorliegende Papier dient folgenden Zielen:

- Es werden ein paar Grundbegriffe eingeführt, deren Kenntnis bei der Zusammenarbeit mit dem Statistiksupport vorausgesetzt wird.
- Es wird das Vorgehen bezüglich der Zusammenarbeit mit dem Statistiksupport erläutert.
- Es werden ein paar allgemeine Richtlinien bezüglich Fragebogengestaltung und Datensammlung erläutert.

Damit soll sichergestellt werden, dass die Zusammenarbeit mit dem Statistiksupport für alle Beteiligten möglichst effizient erfolgt.

1.1 Grundbegriffe

Das Verständnis folgender Grundbegriffe wird bei der Zusammenarbeit mit dem Statistiksupport vorausgesetzt: „Variable“, „Ausprägungen einer Variable“ (= Werte einer Variable); „Kodierung von Variablen“; „Skalenniveaus“ („Nominalskala“, „Ordinalskala“, „metrische Skala“), „univariate, bivariate oder multivariate Hypothese“.

- Eine *Variable* entspricht einer Spalte in einer Datentabelle. Sie teilt eine Grundgesamtmenge von Objekten oder einer Stichprobe auf in Teilmengen (=Äquivalenzklassen), denen wir spezifische Merkmale zuordnen („Ausprägungen“ oder „Werte der Variable“ genannt). Die paarweise Schnittmenge der Äquivalenzklassen müssen leer sein. Die Vereinigungsmenge aller Äquivalenzklassen muss mit der Grundgesamtheit oder mit der Stichprobe identisch sein.
Beispiele:

- Grundmenge Menschen im Wallis. Variable „Geschlecht“ mit den Ausprägungen „männlich“, „weiblich“. Der Ausprägung „männlich“ entspricht die Äquivalenzklasse aller Männer der Grundmenge (oder der Stichprobe), der Ausprägung „weiblich“ die Äquivalenzklasse aller Frauen.
- Grundmenge Schüler von Naters. Variable „Zufriedenheit mit Französischunterricht“ mit den Ausprägungen „sehr zufrieden“, „zufrieden“, „knapp zufrieden“, „unzufrieden“. Zur Äquivalenzklasse „sehr zufrieden“ gehören alle Schüler, die mit dem Unterricht sehr zufrieden sind. etc.
- Grundmenge Schüler der 6. Klasse im Wallis. Variable „Körpergrösse“ mit den Ausprägungen „Körpergrösse in Meter“ (z.B. 1.83, 1.65). Wenn wir genügend genau messen, gehört zu einer Ausprägung im Allgemeinen nur eine Objekt, d.h. jede Äquivalenzklasse enthält nur ein Element.

- Grundmenge Eltern: „liest viel“, „liest nur Zeitung“, „nie“, stellt keine Zerlegung der Eltern bezüglich Leseverhaltens dar, da es Eltern geben kann, die nur Zeitung lesen, die aber viel lesen.
- *Kodierung einer Variable und Skalenniveaus.* Wir kodieren eine Variable, indem wir den Objekten der Grundmenge oder der Stichprobe Zahlen zuordnen (= messen). Die Zuordnung muss gewissen Regeln gehorchen:
 - Objekten derselben Äquivalenzklasse muss jeweils dieselbe Zahl zugeordnet sein. Ist nur diese Regel erfüllt, sprechen wir von „*nominalskalierten Variablen*“ (z.B. Variable Geschlecht, Religionszugehörigkeit, Muttersprache, etc.).
 - Kann man bei den Ausprägungen eine „natürliche“ Ordnung ausmachen, so kann die Ordnung $>$ zwischen den Zahlen die Ordnung zwischen den Ausprägungen spiegeln. Ist dies der Fall und wird (a) erfüllt, so nennen wir die Variable „*ordinalskaliert*“. Sonst bleibt die Zuordnung willkürlich - der Abstand zwischen den Zahlen ist belanglos. (z.B. Zufriedenheit mit Französischunterricht, 4 für „sehr zufrieden“, 3 für „zufrieden“, 2 für „mässig zufrieden“, 1 für „unzufrieden“).
 - Gilt (a) und (b) und spielt der Abstand zwischen den zugeordneten Zahlen eine Rolle (z.B. Längenmessung, Altersmessung), so sprechen wir von einer metrisch skalierten Variable. Wenn Sie über Skalierung mehr erfahren möchten, siehe [Variablen, Ausprägungen und Skalen](http://sozio.logik.ch) auf der Homepage <http://sozio.logik.ch>.
- *Hypothesen* sind das Kernstück statistischer Analysen. Man kan zwei Arten von Hypothesen unterscheiden:
 - Strikte Hypothesen. Es handelt sich um Allsätze wie z.B. Alle Metalle leiten Strom. Alle Menschen sind sterblich. Sie können durch eine einziges Gegenbeispiel falsifiziert werden.
 - Statische Hypothesen sind gesetzesmässige Tendenzauaussagen (z.B. Raucher leben in der Tendenz weniger lang als Nichtraucher. Im Einzelfall kann ein Raucher länger leben als ein Nichtraucher. Statistische Aussagen sagen nicht etwas über alle Raucher und Nicht-raucher einzeln aus, sondern über einen tendenzmässigen, allgemeinen Zusammenhang zwischen Rauchverhalten und Lebensdauer. Entsprechend kann man auf Gebieten, wo nur Tendenzauaussagen möglich sind, nicht mit Einzelfällen argumentieren. Man kann versuchen, mit statischen Hypothesen zu arbeiten, wenn strikte Hypothesen nicht haltbar sind (z.B. die strikte Hypothese „Alle Nichtraucher leben länger als alle Raucher“ ist ein offensichtlich falscher Satz. Wir müssen eine statistische Aussage formulieren: „Im Durchschnitt leben Nichtraucher länger als Raucher“. Entsprechend sind statische Hypothesen in den Humanwissenschaften die Regel und man darf nicht mit Einzelfällen argumentieren: die Tendenzauaussage wird nämlich nicht durch das „Argument“ falsifiziert „Mein Grossvater hat seit 20 jeden Tag 4 Päcklein Parisienne carrée geraucht und wurde 90.“ oder „Meine Cousine hat nie geraucht und starb mit 30 an Lungenkrebs.“). In den Sozialwissenschaften sind statistische Hypothesen die Regel.
- Man unterscheidet zwischen (a) univariaten, (b) bivariaten und (c) multivariaten statischen Hypothesen.
 - Univariate Hypothesen* machen bezüglich einer einzelnen Variablen Aussagen: z.B. Mehr als 50% der PH-Absolventen sind mit der Grundausbildung an der PH zufrieden.
 - Bivariate Hypothesen* behaupten einen tendenzmässigen Zusammenhang zwischen zwei Variablen (z.B. Je mehr Gruppenarbeiten im Französischunterricht, desto lieber mögen die Kinder den Französischunterricht. Je grösser das Einkommen der Eltern, desto höher ist der Anteil der Schüler, die an ein Gymnasium übertreten). Fast alle im Rahmen einer Diplomarbeit zu behandelnden Hypothesen fallen in diese Kategorie. Entsprechend braucht man zur Überprüfung der Hypothese jeweils zwei Variablen.

- c) *Multivariate statische Aussagen* behaupten einen Zusammenhang zwischen mehreren Variablen und einer Zielvariablen (z.B. die Lesekompetenzen hängen von den drei Variablen „Einkommen der Eltern“, „Muttersprache“ und „Geschlecht“ ab). In der Multivariaten Statistik geht es darum, das Gewicht dieser Einflussfaktoren zu schätzen.

1.2 Vorgehensweise beim Kontakt mit der Statistikbetreuung

Bei jedem der folgenden Schritte könnten Sie die Statistikbetreuung um Rat fragen - nach gebührender Vorbereitung des Schrittes. Vor jedem Kontakt sollte dieses Papier genau studiert werden.

1. Sie müssen sich zuerst darüber im Klaren sein, was Sie genau wissen wollen. Allfällige Hypothesen müssen sauber formuliert werden (schriftlich). Hypothesen sind keine Fragen, sondern Aussagesätze, die in der Folge durch die Untersuchung falsifiziert oder bestätigt werden.
2. Stellt man einen Fragebogen zusammen oder sammelt man sonst Daten, muss man wissen, mit welchen Variablen man welche Hypothesen testen will! Erst nachher dürfen Daten gesammelt werden! (Auswertung von Texten, Fragebogen, Beobachtung, etc.). Schreiben Sie auf, mit welchen Daten und mit welchen Variablen Sie die Hypothesen zu überprüfen gedenken. Geben Sie bei den Variablen die Ausprägungen an.
 - Erheben Sie nur Daten, die Sie für die Überprüfung der Hypothesen verwenden (oder für die Sie ein im Rahmen ihrer Arbeit klar beschriebenes Wissensbedürfnis bezüglich beschreibender Statistik haben). Erstellen Sie möglichst kurze Fragebögen.
 - Bei Variablen ist es wichtig, dass durch sie eine *Zerlegung* der Grundmenge oder der Stichprobe festgelegt sein muss: (a) allen untersuchten Objekten muss *vom Konzept her* eine Ausprägung der Variable zuzuordnen sein (in der Wirklichkeit gibt es natürlich immer fehlende Daten). (b) Keinem Objekt sind pro Variable mehr als eine Ausprägung zuzuordnen. Dies hat etwa bei Fragebogen folgende Bedeutung: Mehrfachantworten zu einer Frage sind nur sinnvoll, wenn man die Anzahl der Antworten zählen will. Ist dies nicht der Fall, macht man besser für jede mögliche Antwort eine eigene Variable. In diesem Falle muss man den Fragebogen so gestalten, dass alle Personen alle Fragen beantworten. Bei Mehrfachantworten wird sonst für jede mögliche Antwort eine dichotome Variable geschaffen (dichotom = zwei Werte der Variablen). Dies führt zu einer Vermehrung der Variablen und diese werden, wie die Praxis zeigt - selten bei der Interpretation genutzt. Mehrfachantworten sollte man also nur ermöglichen, wenn man vor der Lancierung des Fragebogens weiß, was man mit den Antworten konkret machen wird!!!!!
 - Testen Sie Fragebögen an ein paar Personen und diskutieren Sie den Fragebogen mit diesen Personen (um Mehrdeutigkeiten, unklare Fragestellungen etc. herauszufinden).
3. Statistische Tests setzen voraus, dass die Daten als Werte von unabhängigen und identisch verteilten Zufallsvariablen betrachtet werden können. Dies ist bei schulischen Untersuchungen selten der Fall. Manchmal versucht man dem Mangel zu begegnen, indem man eine Zufallsstichprobe aus einer Grundgesamtheit zieht: jedes Objekt der Grundgesamtheit hat dabei eine angebbare Wahrscheinlichkeit, in die Stichprobe aufgenommen zu werden. Am besten besorgt man sich eine Liste der Objekte der Grundgesamtheit und entnimmt diesen mit Excel eine Zufallsstichprobe (man ordnet eine Spalte mit „zufallszahl()“ (französisch: alea()) zu und ordnet dann das ganze nach der Spalte mit den Zufallszahlen. Dann werden die ersten n Objekte ausgewählt ($n =$ Stichprobengröße)). Es gibt noch weitere Verfahren, um Zufallsstichproben zu ziehen. Nehmen Sie bei diesbezüglichen Problemen Kontakt mit dem Support auf. Es muss allerdings betont werden, dass solche Verfahren im Allgemeinen für Diplomarbeiten zu schwerfällig sind. Um didaktische Verfahren zu überprüfen, werden diese auf ganze Klassen angewendet. Man müsste also Zufallsstichproben von Klassen, und nicht von Schülern ziehen - eine genügend grosse Menge von Klassen und die Anwendung

verschiedener didaktischer Verfahren auf Teilmengen dieser Klassen ist im Rahme einer Diplomarbeit kaum möglich. Der Zweck der statistischen Verfahren in diesem Zusammenhang ist entsprechend eher, sich mit solchen bekannt zu machen. In der Diplomarbeit muss man auf die entsprechenden Grenzen der Aussagekraft von Tests hinweisen (Eingestandene Sünden sind lässliche Sünden!?!).

4. Bei der Erfassung der Daten werden die Ausprägungen der Variable kodiert, indem für die Ausprägungen Zahlen festgelegt werden (dies ist bei metrisch skalierten Daten oder bei Zähldaten nicht nötig). Bei ordinalskalierten Daten (z.B. sehr gut, mässig gut, unzufrieden) muss die Kodierung so erfolgen, dass der Ordnung der Merkmale Rechnung getragen wird (z.B. „besser“ entspricht „>“ in den Zahlen, so wird bei „sehr gut“, z.B. 4 zugeordnet, bei „mässig gut“ „3“ und bei „unzufrieden“ „2“). Bei nominalskalierten Daten ist die Kodierung beliebig (z.B. bei der Variable „Geschlecht“ wird den Frauen 1 zugeordnet und den Männern 2 - oder umgekehrt). Bei metrisch skalierten Daten erhalten Sie durch Messung unmittelbar die Kodierung selber (z.B. man misst die Körpergrösse oder man zählt die Anzahl Blickkontakte der Lehrperson mit den Schülern).
5. Für die Abgabe der Daten zwecks statistischer Auswertung sollten diese in **einer** Exceldatei erfasst werden, wenn die Daten verschiedener Tabellen verglichen werden sollen. Erheben Sie Daten zweier Grundgesamtheiten, müssen diese immer eine Beziehung haben - wenn Sie diese vergleichen wollen. Sie müssen sich für eine der Grundgesamtheiten entscheiden und die andere in Variablen der ersten ausdrücken (z.B. Sie haben eine Schülerdatentabelle und die Meinungen verschiedener Lehrer. Dann hat jeder Schüler einen Lehrer x mit Meinung y . Wir haben somit die Variable „... hat Lehrer x mit Meinung y “. Wir kodieren die möglichen Antworten x . Hat der Schüler noch einen 2. oder 3. Lehrer, schaffen wir jedesmal eine Variable „hat Lehrer x mit Meinung y “.
6. **Jeder Spalte entspricht eine Variable, jeder Zeile ein Untersuchungsobjekt, jeder Zelle (= Schnittfläche von Zeile und Spalte) entspricht genau ein Datum** (Bitte keine Leerspalten oder Leerzeilen in der eigentlichen Datenliste machen). Siehe Beispiel: **Beispiel Datentabelle**. Nehmen Sie eine möglichst vollständige Aufnahme der gesammelten Daten in die Tabelle vor (man kann nachher leicht umkodieren, Information reduzieren oder Variablen zusammennehmen).
7. Werden an denselben Untersuchungsobjekten zweimal Daten erhoben, so gilt immer noch: jeder Zeile entspricht ein Untersuchungsobjekt. Die Daten der zweiten Erhebung sind als neue Variablen einzufügen.
8. Die Daten müssen zwecks statistischer Auswertung in einer SPSS-Datei an den Statistik-Support geliefert werden, ausser es liegen wenige Variablen vor (dann kann dies der Statistik-Support machen und es genügt eine korrekte Excel-Tabelle abzuliefern). Es empfiehlt sich, die Daten zuerst in einer Excel-Tabelle zu sammeln. Nachdem die Daten vollständig gesammelt sind, kann man sie in eine SPSS-Datei kopieren. SPSS ist ein Statistikprogramm, das verschiedene Editoren aufweist („Ansichten“ genannt). Uns interessieren die Variablenansicht und die Datenansicht. Einer dieser beiden Editoren erscheint, wenn man SPSS startet. Gehen Sie fürs Kopieren auf die Datenansicht, kopieren Sie die Daten in Excel (Ctrl + Shift + c) und wählen Sie dann in der Datenansicht von SPSS „Bearbeiten“, „Einfügen“. Kontrollieren Sie, ob die Daten richtig eingefügt wurden. Als nächstes müssen Sie die Variablen und die Ausprägungen der Variable im Variableneditor korrekt beschriften (möglichst kurze Ausrücke wählen). Dazu gehen Sie auf die Variablenansicht. Geben Sie unter „Variablenlabel“ den Namen der Variable ein (dieser Name erscheint nachher in vom Statistiksupprt zugesandten Tabellen und Graphiken). Unter „Wertelabel“ geben Sie jeweils die Ausprägungen der Variable an (Doppelklick rechts im Feld; auch diese erscheinen nachher in Tabllen und Graphiken. Bitte korrekt schreiben. Nicht zu lange, da sonst nicht alles ausgedruckt wird). Müssen Sie mehrmals dieselben Ausprägungen eingeben, können Sie diese kopieren (mit

Ctrl+Shift, c) und einfügen mit (Ctrl+Shift,v) (Variante: man kann eine Excel-Datei unmittelbar einlesen: Datei, Öffnen, Daten, xls-Format wählen, Datei suchen. Diese Variante ist besonders nützlich, wenn man in die erste Zeile der Excel-Tabelle die Namen der Variablen samt Wertnamen eingegeben hat, da diese dann auch eingelesen wird).

Achtung: Verändern sie in SPSS die erste Spalte in der Variablenansicht nicht. Verändern Sie auch die letzte Spalte mit den Skalenniveaus nicht. Nur die Spalte „Variablenlabels“ und „Wertelabes“ dürfen bearbeitet werden.

9. Bei der Abgabe der Daten beim Statistiksupport sollten nebst
 - (a) der korrekt ausgefüllten SPSS-Tabelle (Excel-Tabelle bei wenigen Variablen) auch
 - (b) eine Liste mit den Hypothesen mitgeliefert werden (samt den Angaben, welche Hypothesen mit welchen Variablen zu testen sind, jeweils in Klammer nach der Hypothese).
 - (c) Wurden die Daten mit einem Fragebogen erhoben, so ist auch der Fragebogen mitzusenden.
 - (d) Eine Beschreibung des Datenerhebungsverfahrens (Angabe der Grundgesamtheit; Grösse der Grundgesamtheit, Grösse der Stichprobe, Wie wurde die Stichprobe erhoben).

Bei korrekter Abgabe der Tabelle und der Variablen schickt der Statistiksupport eine Auswertung in htm-Form zurück (beschreibende Statistik; Teststatistik bezüglich der Hypothesen).

e-mail-Adresse: Paul.Ruppen@hepvs.ch

Bitte geben Sie bei der Zustellung von Daten Ihre Telephonnummer und Anrufzeiten an, damit - falls nötig - zurückgerufen werden kann. . Die Statistik-Betreuung findet im Allgemeinen per e-mail statt, wenn nötig per Telephon, bei Blockaden ausnahmsweise auch vor Ort.

Unter <http://sozio.logik.ch> (Statistiksupport PH Wallis) finden Sie weiter Angaben. (Version 20. September 2021)