Einführung in die Statistik für Digital Humanists - Sommersemester 2022

Hypothesenprüfung in der Statistik

Operationalisierung

Heute wollen wir uns mit der Frage beschäftigen, wie statistische Verfahren eingesetzt werden, um eine Hypothese zu prüfen. Der erste Schritt von der Theorie zur statistischen Prüfung ist die *Operationalisierung*. Was bedeutet dieser Begriff? Nun, stellen wir uns vor, wir wären Anhänger der folgenden beliebten Ansicht:

> William Shakespeare zeichnet sich als Autor durch einen besonders großen Wortschatz aus.

Das ist unsere Annahme. Um diese Annahme mit einem quantitativen Verfahren zu prüfen, müssen wir sie in eine Form bringen, die man in eine mathematische Rechnung, oder in Programmcode übersetzen kann. So, wie sie jetzt da steht, ist sie dafür zu schwammig. Eine mögliche Operationalisierung dieser Annahme wäre, zunächst einmal den Wortschatz in Form der sog. *Type-Token Ratio* (TTR) der bekannten Texte zu messen. Die TTR ist nichts anderes als die Anzahl der gebrauchten Vokabeln in einem Text geteilt durch seine Länge:

$$TTR = \frac{Anzahl\ der\ Vokablen}{Länge\ des\ Textes}$$

Dann brauchen wir Vergleichsdaten. Das ist besonders wichtig für diese Art von Aussage und wird tatsächlich von vielen, die neu in die quantitative Forschungsmethodik einsteigen, zunächst übersehen. In der qualitativen Forschung wird man Shakespeare-Expertin oder Experte, indem man viel von Shakespeare liest. Wenn die Leute dann in die *Digital Humanities* einsteigen wollen, stellen sie ein Corpus aus Shakespeare-Texten zusammen, und wollen daran die oben genannte Frage untersuchen. Wo liegt nun das Problem? Das Problem liegt in der Frage, was eigentlich ein hoher TTR-Wert ist? Hoch verglichen womit? Vermutlich meinen wir: Hoch verglichen mit dem anderer Autoren. Was wir also in so einem Fall auf jeden Fall noch brauchen, ist ein geeignetes Vergleichscorpus, zum Beispiel ein Corpus mit Dramen anderer englischer Autoren aus dem frühen 17. Jahrhundert. Diese Art von Problem scheint recht trivial und dieses Beispiel ist jetzt fiktiv. Aber tatsächlich - und das spreche ich auch aus meiner Beratungserfahrung - ist dieser Schritt für viele, die von der traditionellen Literaturwissenschaft her kommen, oft zunächst nicht naheliegend. Für einen quantitativen Forschungsansatz brauchen wir in vielen Fällen Vergleiche, und das bedeutet oft, daß wir dem, was uns eigentlich weniger interessiert (die anderen Autoren) mindestens genau so viel Aufmerksamkeit schenken müssen, wie dem eigentlichen Forschungsgegenstand (dem Shakespeare'schen Werk).

Kommen wir nun zurück zur Operationalisierung. Wir nehmen also ein Corpus mit Shakespeare-Dramen, und messen die TTR-Werte. Dann nehmen wir ein Corpus mit anderen Dramen, und messen die TTR-Werte. Von beiden Gruppen

könnte man jetzt die Mittelwerte ($\overline{x}$) nehmen, und diese vergleichen. Die Operationalisierung der Behauptung, Shakespeare habe eine besonders großen Wortschatz, könnte also mathematisch formuliert so aussehen:

$$\overline{TTR}_{Shakespeare} > \overline{TTR}_{andere\ Autoren}$$

Natürlich ist das nicht die einzige mögliche Operationalisierung dieser Annahme. Tatsächlich ist die Frage, ob die gewählte Operationalisierung die eigentliche Forschungsfrage präzise genug abbildet häufig ein Streitpunkt in wissenschaftlichen Diskussionen. Sie werden vielleicht mitbekommen haben, daß Virologen und Epidemiologen eifrig darüber diskutieren, ob eine detailliere Studie zur COVID-19-Ausbreitung in einer einzigen Kleinstadt eine geeignete Operationalisierung sein kann, wenn die Forschungsfrage zum Beispiel lautet, wie sehr Schulen und Kindergärten landesweit zur Seuchenausbreitung beitragen können. Insbesondere in den *Digital Humanities* ist die Frage nach der Operationalisierung sehr zentral. In Forschungsdisziplinen mit einer längeren quantitativen Tradition neigt man generell dazu, bei der Formulierung von Forschungsfragen die Operationalisierung mit im Blick zu haben, und diese Tradition gibt es in der Literaturwissenschaft noch nicht. Die Forschungsgegenstände könnten Konzepte sein, die ohnehin umstritten und schwer abzugrenzen sind. Dennoch müssen wir sie für den quantitativen Ansatz in mathematisch fassbare Konzepte übersetzen, und an diesem Punkt finden natürlich kontroverse Diskussionen statt. Es gibt zum Beispiel Forschungsprojekte, die sich mit dem Sprachstil verschiedener Literaturgenres befassen. Wie man solche Genres definiert, festlegt und abgrenzt ist aber in der Literaturwissenschaft durchaus umstritten.

Nullhypothese und Gegenhypothese

Der klassische Weg, eine Forschungsfrage zu operationalisieren und zu beantworten führt in der Statistik über eine sogenannte *Nullhypothese*. Das Ziel dieses Ansatzes ist folgender: Die Forschungsfrage wird in eine Ja/Nein-Frage übersetzt, die man dann anhand bestimmter Kriterien mit Hilfe messbarer Zahlen beantwortet.

Spielen wir das ein einem Beispiel durch: Nehmen wir die Frage nach dem *gender bias* in der Figurenverteilung eines Dramas. Unsere Annahme lautet:

> Männliche (oder weibliche) Figuren sind in dem Stück überrepräsentiert.

Ein erster Schritt zu einer möglichen Operationalisierung dieser Annahme wäre:

> Die Wahrscheinlichkeiten für den Auftritt einer weiblichen Figur und den einer männlichen Figur sind gleich hoch.

Oder:

$$p_{weibliche\ Figur} \neq p_{männliche\ Figur}$$

Wenn wir die Annahme auf diese Art mathematisch formulieren, gibt es nur eine einzige Alternative, wenn die Annahme nicht zutrifft, nämlich:

$$p_{weibliche\ Figur} = p_{männliche\ Figur}$$

Mathematisch gesehen muss eine dieser beiden Formeln zutreffen. Mit der letzten Formel haben wir nun endlich unsere *Nullhypothese* (H_0) formuliert. Was macht H_0 aus? Zunächst einmal ist sie das Gegenteil unserer Vermutung. Wichtig ist aber auch, daß sie davon ausgeht, daß die Zahlen, die wir beobachten rein zufällig und nicht systematisch zustande kommen. Es ist reiner Zufall, daß da auf der Bühne etwas mehr Männer als Frauen herum laufen, das hat nichts damit zu tun, daß der Autor männliche Figuren interessanter fand.

Die andere, sehr wichtige Eigenschaft von H_0 ist, daß sie mathematisch sehr viel greifbarer ist, als die Ungleichung, mit der wir unserer Annahme formuliert haben. Sie ist so greifbar, daß wir ein formales Verfahren konstruieren können, um sie zu prüfen. Wenn wir eine bestimmte Beobachtung haben, zum Beispiel

> Im Drama XY gibt es 7 männliche und 2 weibliche Figuren

und gemäß der Nullhypothese von gleichen Wahrscheinlichkeiten ausgehen,

$$H_0 : p_{weibliche\ Figur} = p_{männliche\ Figur}$$

dann können wir ausrechnen, wie wahrscheinlich diese Beobachtung ist - dazu später mehr - zumindest unter der Annahme, das H_0 zutrifft. Wenn Beobachtung unter der Annahme von H_0 sehr unwahrscheinlich ist, werden wir H_0 selbst als unwahrscheinlich betrachten, und verwerfen.

Das Konzept der Nullhypothese ist derart zentral in der analytischen Statistik, daß die eigentliche Annahme, mit wir als Forscher die Analyse angehen, eigentlich nur noch als Gegenteil der Nullhypothese als *Gegenhypothese*, *Alternativhypothese* oder H_1 bezeichnet wird. Zusammenfassend funktioniert eine statistische Untersuchung also folgendermaßen:

1. Operationalisiere deine Forschungsfrage und konstruiere ein formalisierte Annahme.
2. Finde eine geeignete **Nullhypothese**.
3. Falsifiziere deine Nullhypothese, indem zu zeigst, daß deine Beobachtung sehr unwahrscheinlich ist, wenn die Nullhypothese stimmt.

Und nochmal zur Zusammenfassung: Was macht eine Nullhypothese aus?

1. Sie geht davon aus, daß die Variabilität der Beobachtungen rein zufällig ist.
2. Sie ist mathematisch modellierbar.
3. Sie erlaubt es, die Zufallswahrscheinlichkeit einer bestimmten Beobachtung zu berechnen (oder zu schätzen).

Wie genau eine Nullhypothese mathematisch modelliert wird, ist von Fall zu Fall sehr verschieden. In Statistikbüchern werden Sie zahlreiche sogenannte *Tests* finden, die für verschiedene Forschungsdesigns entwickelt wurden: Den χ^2-Test, den t-Test, den F-Test und den U-Test, um ein paar bekannte zu nennen. Jeder dieser Tests modelliert eine bestimmte Art von Fragestellung und Nullhypothese, der t-Test zum Beispiel die Nullhypothese, daß zwei Zahlenreihen (z.B. die Länge

von 10 Romanen und 10 Novellen) aus Gruppen mit identischen Mittelwerten stammen. (Übersetzung in eine literaturwissenschaftliche Annahme/Frage: Romane und Novellen sind gleich lang/nicht gleich lang.) In den Übungsaufgaben für heute wollen wir versuchen, uns der Nullhypothese für die Frage nach dem *Genderbias* in Dramen anzunähern.

Der *P*-Wert im statistischen Hypothesentest

Genderbias?

Um die Frage nach dem *Genderbias* statistisch zu untersuchen, haben Sie nun eine Nullhypothese formuliert:

> H_0 : Ob die Autorin/der Autor eine männliche oder eine weibliche Figur wählt, ist gleich wahrscheinlich.

oder:

$$H_0 : p_{weibliche\ Figur} = p_{männliche\ Figur}$$

Warum war das die Nullhypose?

1. Sie ist das binäre **Gegenteil unserer Annahme**, daß ein Autor eine Präferenz hat. (Binär meint, er hat eine Präferenz, oder nicht. Wir gehen nicht davon aus, daß irgend etwas dazwischen möglich ist.)
2. Sie unterstellt, daß tatsächlich beobachtbare Unterschiede **rein zufällig** sind und nicht systematisch/kausal durch einen bestimmten Einfluss zu Stande kommen.
3. Eine Realität, in der die Nullhypothese zutrifft, lässt relativ einfach wahrscheinlichkeitstheoretisch modellieren.

Als nächstes haben wir uns die Frage gestellt: Wenn ich nun ein Drama habe mit 4 männlichen Figuren, und nur einer weiblichen, wie wahrscheinlich ist das eigentlich unter Zufallsbedingungen (also wenn H_0 zutrifft). Und wie Wahrscheinlich ist eine Verteilung von 15:5, wenn wir naiv vermuten, daß jeder Figur einzeln "per Münzwurf" (oder wie im richtigen Leben in der Chromosomenlotterie) zufällig ein Geschlecht zugewiesen wurde?

2 Geschlechter hoch 5 Möglichkeiten

Im ersten Fall ("4:1" oder "5 von 4") gibt es $2^5 = 32$ mögliche Kombinationen. Darunter ist nur eine Kombination ohne eine einzige weibliche Figur.

```
[m, m, m, m, m]
```

In 5 Kombinationen kommt nur eine weibliche Figur vor.

```
[w, m, m, m, m]
[m, w, m, m, m]
[m, m, w, m, m]
[m, m, m, w, m]
[m, m, m, m, w]
```

Das macht insgesamt 6 Kombinationen mit mindestens 4 männlichen Figuren. Die Wahrscheinlichkeit, unter Zufallsbedingungen mindestens 4 männliche Figuren zu haben ist damit

$$\frac{6}{32} = 0.1875$$

oder 18.75%. Die Wahrscheinlichkeit für mindestens 4 weibliche Figuren unter Zufallsbedingungen ist identisch, damit wäre die Wahrscheinlichkeit für ein Ungleichgewicht von mindestens 4:1 zugunsten irgendeines Geschlechts 37.5%.

18.75 männlich + 18.75 weiblich = irgendein Geschlecht

Der *P*-Wert

Wie alle Wahrscheinlichkeiten wird auch diese in Formeln gern mit dem Buchstaben *P* bezeichnet, und darum ist in der Statistik und in Diskussionen über Forschungsergebnisse immer wieder die Rede vom *P*-Wert. Je kleiner dieser *P*-Wert in einer Studie, desto stärker das Argument dafür, daß mehr als nur Zufall im Spiel ist, d.h. daß das Medikament wirkt, daß die Lebenserwartung einer bestimmten Personengruppe geringer ist oder daß Autor XY kein großes Interesse an weiblichen Figuren hat. Der *P*-Wert ist wie die Nullhypothese ein zentrales Konzept der analytischen Statistik; eher mittelmäßig intuitiv, nicht unumstritten, aber aber erst mal besser als die Faktensuche nach Bauchgefühl und so weit verbreitet, daß man ihm fast überall begegnet. In unserem Beispiel ist der *P*-Wert

> die Wahrscheinlichkeit, bei 5 Münzwürfen **mindestens** 4 mal das gleiche Ergebnis zu bekommen, wenn beide Seiten exakt gleich wahrscheinlich oben zum liegen kommen.

Allgemeiner formuliert:

> *P*-Wert: Die Wahrscheinlichkeit für die gegebene oder eine noch extremere Beobachtung, wenn H_0 zutrifft.

Wenn diese Wahrscheinlichkeit nun sehr klein ist, dann gehen wir eher davon aus, daß H_0 nicht zutrifft, und das heißt in den meisten Fällen, daß unsere Vermutung stimmt. Kurzum, in der statistischen Hypothesenprüfung **sammeln wir Evidenz für eine Vermutung, indem wir zeigen wie unwahrscheinlich unsere Beobachtungen unter der gegenteiligen Zufallsvermutung wären**.

Ja, das ist zugegebenermaßen nicht der geradlinigste Gedankengang. Aber wenn man ihn erst einmal verdaut hat, bietet er eine Fülle von Möglichkeiten, verschiedene Hypothesentests zu operationalisieren. Unser rein binäres *Genderbias*-Münzwurf-Problem bietet sich dafür sehr leicht an, weil wir die Wahrscheinlichkeit *P* sehr einfach ermitteln können. Wir können sie durch Simulation annähern (in *Python* oder durch sehr geduldiges Münze-werfen) oder sogar mit Mitteln der Oberstufenmathematik (Kombinatorik) direkt ausrechnen, bei kleineren Zahlen sogar einfach durch aufschreiben aller potentiellen Kombinationen ganz plump abzählen.

Statistische Signifikanz

Aber was genau sagt nun ein *P* von 37.5% für einen Hypothesentest aus? Gibt es jetzt einen *Genderbias*? Ist H_0 wahr oder falsch? Zunächst Vorsicht: *P* ist **nicht** die Wahrscheinlichkeit der Nullhypothese! Das ist ein weit verbreitetes

Missverständnis. P sagt etwas über die Beobachtung aus (s.o.). Wie klein muss die Zufallswahrscheinlichkeit unserer Beobachtung sein, damit wir guten Gewissens nicht mehr an Zufall glauben wollen?

Hier hat sich in den meisten Fächern die Konvention etabliert, einfach bei 5% eine Grenze zu ziehen. Ist P kleiner als 5%, wird H_0 abgelehnt, *ergo* die Beobachtung unterstützt unsere Vermutung, *ergo* der Versuch unsere Hypothese zu falsifizieren ist misslungen. Ist $P < 5\%$, dann spricht man von einem *statistisch signifikanten* Ergebnis. Auch diese Begrifflichkeit stiftet gelegentlich Verwirrung: Statistische Signifikanz ist nicht dasselbe wie inhaltliche Bedeutsamkeit. Ein sehr kleiner Unterschied kann, wenn er konstant auftritt und genug Messdaten vorliegen, statistisch signifikant sein. Ob er deswegen auch wichtig ist, kann der P-Wert nicht sagen. Ein Ungleichgeweicht von 4:1 mit seinem $P = 37.5\%$ ist also nach diesem Kriterium *nicht signifikant*, d.h. keine Evidenz für eine systematische Diskriminierung im Auswahlprozess.

Den Grenzwert von 5% bezeichnet man als ***Signifikanzniveau***, und er wird in Formeln mit dem Buchstaben α bezeichnet. Warum ist α nun ausgerechnet 5%? Die Antwort ist so trivial wie unbefriedigend: Weil die meisten Leute in der Wissenschaft ihn so akzeptieren. Im Prinzip ist dieser Wert willkürlich, und eigentlich spricht überhaupt nichts dagegen in eine Studie zu schreiben, daß man seine Hypothesen mit einem Signifikanzniveau von $\alpha = 0.1\%$ testen möchte. Aber mit 5% ist halt die Mehrheit des Berufsstandes erst mal glücklich.

Der Binomialtest

Wie schon früher angedeutet gibt es eine Reihe von verbreiteten Verfahren, die für eine bestimmte Art von Daten und eine bestimmte Nullhypothese P-Werte berechnen oder abschätzen. Solche Verfahren werden *statistische Tests* genannt, und sie füllen viele Bücher. Einen ersten dieser Tests haben wir hier schon durchgeführt, den sog. *Binomialtest*. Der Binomialtest berechnet einen P-Wert für die Verteilung von Beobachtungen auf genau zwei Kategorien ("weiblich/männlich", "Kopf/Zahl", "0/1" ...) bei einer vorgegebenen Zufallswahrscheinlichkeit für ein Einzelereignis (in unseren Beispielen 50%). Das schöne am Binomialtest ist, daß sich bei kleinen Zahlen sehr leicht nachvollziehen lässt, wie der P-Wert zu Stande kommt. Man kann das natürlich auch in einer Formel ausdrücken. Wenn p die Wahrscheinlichkeit für das einzelne Ereignis ist (1x Kopf bei einem Münzwurf, hier 0.5), k die Häufigkeit dieses Ereignisses (wie oft war "Kopf" oben) und n die Zahl der Beobachtungen (wie oft wurde die Münze geworfen), dann gilt:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

In `Python` ist das alles aber glücklicherweise schon mit dem Paket `scipy` implementiert. Wir können hier also ganz schnell unser Ergebnis reproduzieren:

```
from scipy import stats
stats.binom_test(4, n=5, p=0.5, alternative='two-sided')
```

Aufgabenpaket 5

Abgabe bis Mittwoch 08.06.2022 um 23:59 Uhr

5.1 Testaufgabe:

Nach der Lektüre eines neuen Romans der Autorin X haben Sie den Eindruck, daß die Autorin in diesem Werk mehr wörtliche Rede einsetzt, als in ihren früheren Romanen. Nehmen wir an, Sie haben eine Software, mit der Sie zählen können, wie viele Wörter im Buch wörtliche Rede sind und wie viele nicht.

Formulieren Sie Nullhypothese und Gegenhypothese für eine Operationalisierung Ihrer Annahme.

5.2 Übungsaufgabe:

Kommen wir zurück zum *Genderbias*-Problem. Wenn wir einen *Bias* annehmen, wäre das Gegenteil unserer Annahme, daß verbliebene Ungleichgewichte zufällig zu Stande kommen. Unsere Nullhypothese

$$H_0 : p_{weibliche\ Figur} = p_{männliche\ Figur}$$

wird also exakt durch das Beispiel mit der geworfenen Münze abgebildet: Jedes mal, wenn der Autor eine neue Figur einführt, ist die Wahrscheinlichkeit für jede der beiden Kategorien gleich hoch.

(A) Wir gehen von einem Drama mit 5 Figuren aus. Schreiben Sie alle potentiellen Kombinationen auf, die auftreten können, wenn Sie nacheinander für jede dieser Figuren die Genderzuordnung per Zufall (=Münzwurf) bestimmen. In wie vielen dieser Kombinationen kommt eine Kategorie (männlich oder weiblich) nur 1 mal oder weniger vor?

(B) Wie hoch ist die Wahrscheinlichkeit für ein Genderungleichgewicht von **mindestens** 4:1 (ergo 4:1 oder 5:0) wenn wir davon ausgehen (H_0), daß beide Zuordnungen eigentlich gleich wahrscheinlich sind?

(C) Sie können die geworfene Münze auch in Python mit einem Zufallsgenerator simulieren. Laden Sie das Paket *random* und nutzen Sie die Funktion *random.sample()*, um gemäß H_0 eine hypothetische Genderverteilung für ein Drama mit 20 Figuren zu erzeugen.

(D) Bauen Sie eine *for*-Schleife, um 1000 Genderverteilungen über 20 Figuren zu erzeugen. Berechnen sie für jedes dieser "Dramen" das Genderverhältnis, am besten als Prozentwert, und speichern Sie alle 1000 Ergebnisse in einer Liste, oder besser noch in einem Numpy-Array oder einer Pandas-Series. Können Sie anhand der Quantile dieser Datenreihe sagen, wie viele ihrer Durchläufe ein Ungleichgewicht von **mindestens** 15:5 haben?

5.3 Testaufgabe:

(A) Ermitteln Sie mit dem Binomialtest, ob es unter den Figuren in Ihrem Drama einen statistisch signifikanten *Genderbias* gibt.

(B) Ermitteln Sie mit dem Binomialtest, ob es unter den Autorinnen und Autoren, zu denen Sie Lebensdaten gesammelt haben, einen statistisch signifikanten *Genderbias* gibt.

(C) Wie viele Autoren muss Ihr Datensatz mindestens haben, damit der P-Wert überhaupt ein Signifikanzniveau von 5% erreichen kann? Stellen Sie sich vor, Sie werfen einen Münze und bekommen immer das gleiche Ergebnis. ab dem wievielten Wurf ist die Beobachtung statistisch signifikant?

5.4 Übungsaufgabe:

Wir haben nun eine Prozedur kennen gelernt, um Hypothesen zu prüfen.

1. Annahme operationalisieren.
2. Nullhypothese finden.
3. Mit Hilfe eines Signifikanztests entscheiden, ob die Nullhypothese verworfen werden sollte oder die ursprüngliche Annahme.

Nehmen wir mal an, sie haben in allen drei Schritten alles richtig gemacht und sind zu einer Entscheidung (H_0 oder H_1) gelangt. Die Ergebnisse Ihrer Studie sind klar und eindeutig.

(A) Kann es trotzdem sein, daß Sie sich irren, auch wenn ihre Studie Handwerklich einwandfrei ist? Welche Fehler können immer noch passieren?

(B) Gibt es Möglichkeiten, das Risiko der unter (A) genannten Fehler zu reduzieren?

(C) Können wir die Wahrscheinlichkeit solcher Fehler irgendwie abschätzen?