Statistische Inferenz | 01 | Fisher-Exakt-Test
Musterlösung

Prof. Dr. Roland Schäfer | Germanistische Linguistik FSU Jena

7. November 2024

1 Fisher-Exakt-Test und Stichprobengröße

1. Rekapitulieren Sie die Berechnung des klassischen *Tea-Tasting Lady*-Experiments für 6 richtige Tassen bei 8 Tassen insgesamt (also 3 richtige und ein falscher „Tee zuerst"-Tipp von insgesamt 4 möglichen richtigen „Tee zuerst"-Tipps) auf den Folien.
2. Berechnen Sie die Wahrscheinlichkeiten bzw. den p-Wert für dasselbe Verhältnis von richtigen Tassen, aber bei einer zehn Mal größeren Stichprobe, also 60 Tassen korrekt vorhergesagt bei 80 Tassen insgesamt.
3. Interpretieren Sie das Ergebnis.
4. Stellen Sie die Anfangswerte dieser Berechnung als Vier-Felder-Tabelle dar.

Lösung zu 2

$$p(30\text{ Richtige}) = \frac{\binom{40}{30}\binom{40}{10}}{\binom{80}{40}} = \frac{\frac{40!}{30!(40-30)!}\frac{40!}{10!(40-10)!}}{\frac{80!}{40!(80-40)!}} = \frac{\frac{40!}{30!10!}\frac{40!}{10!30!}}{\frac{80!}{40!40!}} = \frac{(\frac{40!}{30!10!})^2}{\frac{80!}{40!^2}} \approx \frac{(\frac{8{,}16\cdot10^{47}}{9{,}63\cdot10^{38}})^2}{\frac{7{,}16\cdot10^{118}}{6{,}66\cdot10^{95}}} \approx \frac{7{,}19\cdot10^{17}}{1{,}08\cdot10^{23}} \approx 6{,}68\cdot10^{-6} \approx 0{,}00000668$$

Lösung zu 3

Das ist vor dem Experiment die Wahrscheinlichkeit gewesen, durch Raten **genau 30 Richtige** zu erhalten. Zusätzliche Überlegung: Ist das wirklich das, was uns interessiert? Eigentlich interessiert uns für unsere Schlussfolgerung doch eher, wie wahrscheinlich es war, **mindestens ein so gutes Ergebnis** zu erzielen. Das ist das, was der tatsächliche Fisher-Test typischerweise berechnet, und das ergibt in diesem Fall $p \approx 7{,}44\cdot10^{-6}$. Man kann sich das herleiten als:

$$p(30\text{ oder mehr Richtige}) = \frac{\binom{40}{30}\binom{40}{10}}{\binom{80}{40}} + \frac{\binom{40}{31}\binom{40}{9}}{\binom{80}{40}} + \frac{\binom{40}{32}\binom{40}{8}}{\binom{80}{40}} + \cdots + \frac{\binom{40}{40}\binom{40}{0}}{\binom{80}{40}} \approx 7{,}44\cdot10^{-6}$$

Es fällt auf, dass mit steigender Stichprobengröße trotz einer gleichen Erkennungsrate für die relevanten Teetassen der p-Wert kleiner wird. Es ist also unwahrscheinlicher, 30 von 40 richtig zu raten, als 3 von 4 (wenn man die relevante Fähigkeit nicht hat). Das sollte hoffentlich intuitiv auch angemessen sein, gerade wenn die Fähigkeit der Tea-Tasting Lady, die Reihenfolge des Einschenkens zu erkennen, nicht absolut ist, aber trotzdem eine echte sensorische Fähigkeit darstellt. Technisch gesprochen hängt der p-Wert von der **Effektstärke** (der Qualität der wahren Fähigkeit der Dame, die Reihenfolge des Einschenkens zu Erkennen) ab. Die Wahrscheinlichkeit, einen niedrigen p-Wert zu erhalten, wenn die Tea Tasing-Lady die relevante Fähigkeit wirklich hat, hängt außerdem von der **Größe der Stichprobe** (und damit bei diskreten Ereignissen auch der Größe des Ereignisraums) ab. Je Größer die Stichprobe ist, desto größer wird die Chance, die Fähigkeit zu erkennen.

Lösung zu 4

30	10
10	30

2 Vorgriff auf Kollostruktionsanalysen mit Fisher-Test

Uns werden im weiteren Verlauf sogenannte *Kollostruktionsanalysen* begegnen. Diese sind ein Analyseverfahren für Korpusdaten und wurden ursprünglich (bei ihrer Erfindung vor 20 Jahren) mit Fisher-Exakt-Tests gerechnet. Ein möglicher Datensatz wäre folgender:[1]

	kaufen	***umarmen***
im Passiv	120	30
im Aktiv	380	470

Konzeptionell sind die tabulierten Zahlen die Anzahlen der Vorkommnisse von (in diesem konkreten Fall) dem Verb *kaufen* im Aktiv und im Passiv und dem Verb *umarmen* – ebenfalls im Aktiv und im Passiv – in einer Stichprobe aus irgendeinem Korpus (= Textsammlung) des Deutschen. (Die Zahlen sind hier der Einfachheit halber ausgedacht.) Wir möchten mit so einer Untersuchung quantifizieren, wie hoch **die Affinität der Verben (im Vergleich zueinander) zur Passivbildung** ist.

1. Wie viele Vorkommnisse von *kaufen* und *umarmen* haben wir jeweils beobachtet?
   Je 500.
2. Wie viele Sätze im Passiv und im Aktiv haben wir jeweils beobachtet?
   Im Passiv 150, im Aktiv 850.
3. Wie groß war die gesamte Stichprobe?
   Genau 1000.
4. Was ist das Verhältnis zwischen Passiv und Aktiv bei *kaufen* und bei *umarmen*?
   kaufen: $\frac{120}{380} \approx 0,32$ *umarmen*: $\frac{30}{470} \approx 0,06$
5. Wie würden Sie den p-Wert berechnen? Es reicht der Lösungsweg, wenn Ihr Taschenrechner bei den großen Zahlen aussteigt.
   $p(?) = \frac{\binom{500}{120}\binom{500}{30}}{\binom{1000}{500}} \approx 1,03 \cdot 10^{-133}$
6. Der p-Wert ist ungefähr $1,03 \cdot 10^{-133}$. Überführen Sie diese Zahl aus der Exponentialschreibweise in herkömmliche Fließkommanotation, falls Sie das noch können.
   $0,00000000000000000000000000000000000000000000000000000000000000000000000000000000$
   $0000000000000000000000000000000000000000000000000000138$. Das sind 132 Nullen zwischen dem Komma und der ersten Stelle der Basis.
7. Was sagt uns dieser p-Wert? Was ist das für eine Wahrscheinlichkeit? Das ist ja zunächst mal von der Idee her etwas völlig anderes als bei der *Tea-Tasting Lady*.
   Das war **vor der Korpusstudie** die Wahrscheinlichkeit, **genau** 120 Passive von *kaufen* (bei insgesamt 500 Belegen für *kaufen*) und 30 Passive für *umarmen* (bei insgesamt 500 Belegen insgesamt für *umarmen*) zu finden, **falls der Anteil der Passive bei beiden Verben in der Realität gleiche ist.** Der empirisch interessantere p-Wert für „eine so starke oder stärkere Abweichung" liegt bei $2,59 \cdot 10^{-16}$.

[1] Normalerweise nimmt man etwas andere Designs, aber das schauen wir uns dann in Ruhe an.

8. Es gibt zwischen den Designs der *Tea-Tasting Lady* und der Kollostruktionsanalyse einen wesentlichen Unterschied bezüglich der **Summen der Werte in den Spalten und den Zeilen der Tabelle**. Finden Sie den? Das ist allerdings eine optionale Transferaufgabe auf sehr hohem Niveau.
   Bei der Tea-Tasting Lady sind die Summen der Werte in den Zeilen und Spalten der Tabelle durch das Design des Experiments festgelegt. Es wurde vereinbart, dass sie 8 Tassen bekommt, von denen in 4 die Milch zuerst eingeschenkt wurde. Außerdem sucht sie genau 4 Tassen aus. Egal, wie gut sie rät oder die Tassen erkennt, in jeder Zeile und Spalte der Tabelle ist die Summe der Werte 4. In der fiktiven Korpusstudie haben wir zwar für beide Verben 500 Belege gezogen, und die Spalten summieren sich daher jeweils zu 500, aber wie viele Passive und Aktive wir jeweils finden würden, konnten wir vor der Durchführung der Studie nicht wissen. Die erste Zeile summiert sich zu 150, die zweite zu 850, es hätte aber auch ganz anders kommen können. Daher ist der Fisher-Test eigentlich nicht geeignet für solche Studien. Überlegen Sie, warum. Das ist aber wirklich extrem fortgeschritten. Kaum jemand in der Linguistik weiß das überhaupt, ganz zu schweigen davon, zu wissen, wie es mathematisch zu begründen ist.