Stochastik für die Naturwissenschaften

Dr. C.J. Luchsinger

9. Testen von Hypothesen

Literatur Kapitel 9

* Statistik in Cartoons: Kapitel 8
* Stahel: Kapitel 8 und 10
* Storrer: Kapitel 44-47

Ziele dieses Kapitels:

Das gesamte Setting (die Philosophie, 9.1) rund um das Testen muss verstanden werden. Ausgehend von den in der Vorlesung behandelten Situationen sollten Sie auch fähig sein, Aussagen über statistische Tests in komplizierteren Modellen, welche Sie (noch) nicht kennen, nachzuvollziehen. Zuerst ein paar Vorbemerkungen:

9.1 Philosophie hinter den Tests / Problemstellung

9.1.1 Philosophie hinter den Tests

Frage an's Publikum: (abgesehen von ethischen Fragen), was ist die Grenze der akademischen Forschungsfreiheit? Welche Theorien dürfen Sie vertreten und welche nicht?

* Karl Raimund Popper (1902-1994); ausserhalb der Naturwissenschaften umstritten bis verhasst

* unbekannte Naturgesetze; gehen davon aus, dass sie existieren, konstant in Raum und Zeit und unabhängig von Beobachter

* Gravitation:

"Jeder Stein fällt nach unten" > nicht *verifizierbar*

"Jeder Stein fällt nach oben" > *falsifizierbar*

* vorläufige Resultate, solange nicht widerlegt

* wir kennen nur Theorien und die Resultate von Experimenten und Beobachtungen

* Beobachtungen > Theorie > Experiment zur Überprüfung > entweder bestätigt (aber nicht bewiesen) oder falsifiziert

* Falsifizierung bedeutet Fortschritt in der Forschung

* Die Existenz einer übersinnlichen Gottheit oder die Frage eines freien Willens sind mit obigem Setting wohl nicht entscheidbar

* Akzeptieren Sie Ihr Unwissen und diese Unsicherheit mit Demut; im Gegensatz zu Ingenieuren und Mechanikern, welche eine Maschine bauen und genau wissen, was sie warum dort einbauen, wissen wir nicht, was alles von der Biochemie im menschlichen Körper oder von der Physik in der Atmosphäre am wirken ist (und das *warum* macht eventuell kaum Sinn)

* Mathematik einzige Wissenschaft, in der man etwas wirklich sicher beweisen kann

* alles andere unsicher; arbeiten mit Hypothesen

* Hypothesen müssen falsifizierbar sein (sonst ist es Meta-Physik)

* bis jetzt nur weisse Schwäne gesehen ⇒ Theorie (=Hypothese): "alle Schwäne sind weiss"

* falsch, sobald ein schwarzer Schwan auftritt (falsifizierbar)

* Also: "alle Schwäne sind weiss" ist eine Hypothese, mit der wir arbeiten, bis sie widerlegt wurde (danach eventuell verfeinert weiterarbeiten).

* Popper lieferte Wissenschaftstheorie, "wie sollen wir forschen?", zumindest in den Naturwissenschaften (Buchtitel: Logik der Forschung)

9.1.2 Problemstellung

Stellen wir uns vor, wir sind in einer Quizshow: Gegeben sind zwei *mögliche* Hypothesen, wir nennen sie $\mathcal{H}_0$ und $\mathcal{H}_1$. Entweder ist die Hypothese $\mathcal{H}_0$ richtig oder die Hypothese $\mathcal{H}_1$. Wir wissen leider nicht, welche Hypothese wirklich richtig ist. Wir erhalten dann eine Realisation x_1 einer Zufallsgrösse. Je nachdem, welche Hypothese richtig ist, wird die Zufallsgrösse aber anders verteilt sein (z.B. anderer Mittelwert). Wir müssen uns dann entscheiden, ob wir sagen wollen: "wir nehmen **vorläufig / provisorisch** $\mathcal{H}_0$ an" oder "wir nehmen **vorläufig / provisorisch** $\mathcal{H}_1$ an". 4 Situationen sind möglich:

1. $\mathcal{H}_0$ ist richtig und wir nehmen **vorläufig / provisorisch** $\mathcal{H}_0$ an.
2. $\mathcal{H}_0$ ist richtig und wir nehmen **vorläufig / provisorisch** $\mathcal{H}_1$ an (Fehler 1. Art).
3. $\mathcal{H}_1$ ist richtig und wir nehmen **vorläufig / provisorisch** $\mathcal{H}_1$ an.
4. $\mathcal{H}_1$ ist richtig und wir nehmen **vorläufig / provisorisch** $\mathcal{H}_0$ an (Fehler 2. Art).

Beispiele für $\mathcal{H}_i$ sind:

* $\mathcal{H}_0$: Würfel fair (1/6) gegen $\mathcal{H}_1$: Würfel verfälscht (viele Möglichkeiten)
* $\mathcal{H}_0$: Münze fair (1/2) gegen $\mathcal{H}_1$: Münze verfälscht (viele Möglichkeiten)
* $\mathcal{H}_0$: Medikament lässt Blutdruck unverändert gegen $\mathcal{H}_1$: Medikament ändert Blutdruck
* (Lohn-)Umfrage unter 10'000 Personen, schliesse auf Gesamtpopulation: $\mathcal{H}_0$: Durchschnittslohn gleich geblieben gegen $\mathcal{H}_1$: Durchschnittslohn anders

Erste Aufgabe Gegeben sei eine Realisation x_1 einer Zufallsgrösse aus einer $U[a, b]$-Verteilung. Im Fall $\mathcal{H}_0$ ist dies eine $U[0, 1]$-Verteilung, im Fall $\mathcal{H}_1$ ist es eine $U[2, 3]$-Verteilung. Wir müssen uns jetzt mit Beobachtung x_1 entscheiden, ob wir im Fall $\mathcal{H}_0$ oder $\mathcal{H}_1$ sind:

Zweite Aufgabe (Ziel: **Denken in Hypothesen**) Gegeben sei eine Realisation x_1 einer Zufallsgrösse aus einer $\mathcal{N}(\mu, \sigma^2)$-Verteilung. $\sigma^2 = 1$ gelte auf jeden Fall; aber: im Fall $\mathcal{H}_0$ ist dies eine $\mathcal{N}(0, 1)$-Verteilung, im Fall $\mathcal{H}_1$ ist es eine $\mathcal{N}(1, 1)$-Verteilung. Wir müssen uns jetzt in dieser Quizshow mit dieser einzigen Beobachtung x_1 entscheiden, ob wir im Fall $\mathcal{H}_0$ oder $\mathcal{H}_1$ sind:

Dritte Aufgabe Wir behandeln nochmals die zweite Aufgabe mit folgenden Auflagen:

a) Wenn wir uns für $\mathcal{H}_1$ entscheiden, obschon $\mathcal{H}_0$ richtig ist, verlieren wir all unser Vermögen (in den 3 anderen Möglichkeiten geschieht nichts). Wie sollten wir uns jetzt entscheiden?

b) Wenn wir uns für $\mathcal{H}_0$ entscheiden, obschon $\mathcal{H}_1$ richtig ist, verlieren wir all unser Vermögen (in den 3 anderen Möglichkeiten geschieht nichts). Wie sollten wir uns jetzt entscheiden?

Vierte Aufgabe Wir behandeln nochmals die zweite Aufgabe mit folgender Auflage: Wenn $\mathcal{H}_0$ richtig ist, dürfen wir (ohne Strafe) mit 10 % Wahrscheinlichkeit eine falsche Entscheidung treffen (sagen: "$\mathcal{H}_1$ ist richtig", Fehler 1. Art). Wie sollten wir uns jetzt entscheiden?

Alternative Vorschläge:

Fünfte Aufgabe Wir wollen die vierte Aufgabe zu einer Optimierungsaufgabe umformulieren. Wir haben gesehen, dass wir ∞-viele Möglichkeiten haben, diese 10 % der Fälle "auf der x-Achse zu platzieren" (= sog. Ablehnungsbereich (engl. Critical Region), wir lehnen dort die $\mathcal{H}_0$-Hypothese ab). Wir sollten deshalb noch den Fehler 2. Art mit einbeziehen. Ziel ist es, "**pro Fehler 1. Art, den man macht, ein Maximum an Fehler 2. Art vernichten!**" Wie sollten wir uns jetzt entscheiden?

Wichtige Warnung: Folgende Aussage ist (obschon häufig gehört) schwachsinnig: "Mit 90 % Wahrscheinlichkeit ist die Hypothese $\mathcal{H}_0$ die richtige." Richtig ist: Entweder ist $\mathcal{H}_0$ die richtige Hypothese oder $\mathcal{H}_1$. Wir werden aber nie (Ausnahmen sind triviale Fälle wie die erste Aufgabe) wirklich wissen, in welcher Situation wir sind. Wir werden uns in 10 % der Fälle, wo eigentlich $\mathcal{H}_0$ richtig wäre, für $\mathcal{H}_1$ entscheiden, weil uns dies erlaubt wurde. Wir machen dies aber möglichst sinnvoll (vgl. fünfte Aufgabe).

Die bisherigen Aufgaben waren sehr künstlich, weil wir nur $n = 1$ gehabt haben. In der Praxis haben wir aber hoffentlich ein grosses n.

Aufgabe aus dem Intro (Kapitel 0):

Wenn wir eine Münze (oder einen Würfel) 1'000 mal werfen und die Anzahl Kopf (beim Münzwurf) zählen, so *erwarten* wir eine Zahl wie 500. ... Wenn wir uns fragen, welche Werte dabei etwa so vorkommen, so kommen neben der Zahl 500 auch Zahlen wie 503, 488, 509, 511, 514, 478, und so weiter in den Sinn. Irgendwann wird wohl klar, dass theoretisch eigentlich jede ganze Zahl aus der Menge

$$\{0, 1, 2, \ldots, 998, 999, 1000\}$$

als Resultat vorkommen könnte. Wir werden nicht stutzig, wenn jemand sagt, er habe die Münze 1'000 mal geworfen und dabei genau 512 mal Kopf erhalten. Wenn jemand aber kommt und sagt, ich habe genau 377 mal Kopf erhalten, so fragen wir uns, ob die Münze fair ist oder nicht. Nun kann man sich fragen, wo wir die Grenze ziehen sollten: wo sollten wir Bedenken anmelden ob die Münze auch wirklich fair ist oder nicht. Bei 400 (und 600 gegen oben (Symmetrie)), bei 450 (550), 490 (510)? Kann man so etwas überhaupt für alle Situationen und Personen verbindlich festschreiben? Dieses Problem lässt sich kurz in einer trivialen Feststellung so zusammenfassen und ist der Grund dafür, dass wir überhaupt durch diese Vorlesung müssen:

Es gibt Variabilität in den Daten!

Genauer: *Selbst bei einer fairen Münze* wird es um die 500 herum schwanken. Wie weit darf es aber schwanken, bis wir stutzig werden (sollten)?

Rechnung mit Hilfe des CLT:

Eine solche Testsituation nennt man *zweiseitig*.

Jargon und Bemerkungen:

1. In obigem Beispiel nennt man $\sum_{i=1}^{1000} x_i$ (alternativ $\overline{x}$) die **Teststatistik**; das ganze ist ein **statistischer Test**.

2. α ist Wahrscheinlichkeit für Fehler 1. Art (auch Risiko 1. Art oder Grösse des Tests oder Signifikanzniveau, engl. Size). Bei Aufgabe 4 und 5 hatten wir $\alpha = 10\%$; beim Münzwurf $\alpha = 5\%$.

3. β ist Wahrscheinlichkeit für Fehler 2. Art (auch Risiko 2. Art; Macht ist $1-\beta$; englisch: Power).

4. Signalwörter für $\mathcal{H}_0$ sind: bisherige Stand des Wissens; gerecht/fair $(1/2, 1/6)$, symmetrisch, unspektakulär/langweilig.

5. Bemerkungen zu einseitig vs zweiseitig am Beispiel des Münzwurfs:

Zweiseitig: Wegen 4. werden wir üblicherweise $\mathcal{H}_0$ so wählen, dass dort $p = 0.5$ (symmetrisch, langweilig, gerecht/fair). $\mathcal{H}_1$ ist dann $p \neq 0.5$. Das ist dann eine zweiseitige Testsituation. Signalwörter sind "hat Mondstand einen *Einfluss* auf die Wahrscheinlichkeit" oder "gibt es wegen des Mondstandes eine *Änderung* der Wahrscheinlichkeit"? Ein Einfluss bzw eine Änderung kann ja auf beide Richtungen sein.

Einseitig: Wenn ich aber in einem Glücksspiel bin (und wenig "Kopf" will), so ist mir eine Verfälschung zu meinen Gunsten egal. Dann kann es Sinn machen, zu wählen: $\mathcal{H}_0$ heisst $p \leq 0.5$ und $\mathcal{H}_1$ ist dann $p > 0.5$. Das ist dann eine einseitige Testsituation. Signalwörter sind Steigerung/Erhöhung bzw Senkung. Hier gibt es dann 2 Möglichkeiten (gegen links oder gegen rechts).

6. Wenn die Resultate der Teststatistik derart sind, dass wir $\mathcal{H}_0$ verwerfen, sagt man, dass die Resultate *signifikant* sind (zB signifikant von 0 verschieden).

7. Bemerkungen zu α: Wir haben bis jetzt $\alpha = 0.1$ oder $\alpha = 0.05$ gewählt. Warum nicht (aus Fairness-Gründen) $\alpha = 0.5$ oder besser: Risiko 1. Art = Risiko 2. Art ("Zweite Aufgabe")? In der Praxis hat man meist eine Situation derart, dass $\mathcal{H}_0$ den bisherigen Stand des Wissens repräsentiert. Wir sind sehr konservativ in dem Sinne, dass wir nicht vorschnell zufällige Resultate in gefestigte Lehrmeinungen überführen wollen. Das heisst, wir wählen eher ein kleines α. Genauer gilt folgende Dynamik:

* α gross: nehmen schneller $\mathcal{H}_1$ an (gut, wenn $\mathcal{H}_1$ richtig / schlecht, wenn $\mathcal{H}_0$ richtig)

* α klein: nehmen weniger oft $\mathcal{H}_1$ an (gut, wenn $\mathcal{H}_0$ richtig / schlecht, wenn $\mathcal{H}_1$ richtig)

Da Sie als Forschende spektakuläre Resultate präsentieren wollen (und nicht langweilige / bisherige Stand des Wissens), haben Sie die Neigung, ein eher grosses α zu favorisieren. Das Journal oder die finanzierende Behörde wollen aber keine Schlappe/Skandal (mit sich nachträglich als falsch herausstellenden Resultaten; das gefährdet den guten Ruf) und wollen eher ein kleines α. Schlechter Stil ist dann, nachträglich das α so zu wählen, dass die Resultate signifikant werden (zB α von 1 auf 5 Prozent erhöhen). Wir üben in Vlsg und Ue mit $10, 5, 2, 1\%$.

8. P-Wert:

* kann *vor* dem Versuch im Büro bei Luchsinger nicht berechnet werden; braucht Daten! Damit wird er von Versuch zu Versuch (von Student zu Student) anders ausfallen (wichtig für Leute, welche die R-Aufgaben einfach kopieren).

* Anschaulich ist der P-Wert die Wahrscheinlichkeit unter $\mathcal{H}_0$, dass ein so extremer oder noch extremerer Wert der Teststatistik vorkommt. Der P-Wert ist damit ein Mass dafür, wie signifikant das Resultat ist.

* α ist offenbar eine schöne Zahl; der P-Wert normalerweise nicht.

* Wegen Punkt 5 machen wir *drei* Beispiele mit $n = 1$, $\mathcal{N}(\mu, 1)$ und $\alpha = 5\%$.

a) $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = 1$ und einziger Wert ergab $x_1 = 2.2$:
P-Wert ist dann $P[X_1 \geq 2.2] =$

b) $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = -1$ und einziger Wert ergab $x_1 = -3.1$:
P-Wert ist dann $P[X_1 \leq -3.1] =$

c) $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu \neq 0$ und einziger Wert ergab $x_1 = 0.7$:
P-Wert ist dann $P[|X_1| \geq 0.7] =$

* P-Wert $\leq \alpha : \mathcal{H}_1$ annehmen (Wert Teststatistik eher unwahrscheinlich unter $\mathcal{H}_0$): a), b).
* P-Wert $> \alpha : \mathcal{H}_0$ annehmen (Wert Teststatistik gewöhnlich unter $\mathcal{H}_0$): c).

Kochrezept zum Testen von Hypothesen

Um die Mathematik simpel zu halten, wählen wir $\sigma^2 = 1$ bekannt in einem Beispiel der Normalverteilung.

---

1. Hypothesen aufstellen
2. α und (wenn möglich) n wählen
3. gute Teststatistik wählen
4. Verteilung der Teststatistik unter $\mathcal{H}_0$?
5. Kritischer Wert a
6. Brauchen Daten
7. $\mathcal{H}_0$ annehmen oder ablehnen

Beispiel

---

1. $\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = 1$
2. $\alpha = 0.05, n$ bekannt
3. $\overline{X}$
4. $\mathcal{N}(0, \frac{1}{n})$
5. $P[\overline{X} > a] = P\left[\frac{\overline{X}}{\frac{1}{\sqrt{n}}} > \frac{a}{\frac{1}{\sqrt{n}}}\right] = P[\sqrt{n}\overline{X} > \sqrt{n}a] = P[\mathcal{N}(0,1) > \sqrt{n}a] = 0.05$
6. $\overline{x}$
7. $\mathcal{H}_0$ annehmen wenn $\overline{x} < a$, sonst ablehnen

Alternativ ab Schritt 5: P-Wert mit vorhandenen Daten berechnen und mit gängigen α vergleichen.

Beispiel mit konkretem $n = 12$ und $\overline{x} = 0.67$. Was glauben Sie, werden wir bei $\alpha = 0.05$ die Null-Hypothese verwerfen?

Wir variieren noch n; was ist zu beobachten? Was wenn n seeeeeeeeeeeeeehr gross?

Gute Übung auch wegen Ablesen von Wahrscheinlichkeiten: $\mathcal{H}_0$ sei $\mathcal{N}(0,1)$, $n = 1$, $\alpha = 0.05$. Testen gegen (Test angeben)

a) $\mathcal{H}_1$ ist $\mathcal{N}(1,1)$, bitte auch β berechnen.

b) $\mathcal{H}_1$ ist $\mathcal{N}(2,1)$, bitte auch β berechnen.

c) $\mathcal{H}_1$ ist $\mathcal{N}(3,1)$, bitte auch β berechnen.

d) $\mathcal{H}_1$ ist $\mathcal{N}(4,1)$, bitte auch β berechnen.

e) Zusammenfassung a)-d). Macht es Sinn?

Lesen Sie jetzt Kapitel 44; es folgen auf den weiteren Seiten mehrere bekannte Testsituationen (Storrer Kapitel 45-47):

Im Kochrezept zum Testen von Hypothesen heisst es in Punkt 3: "gute Teststatistik wählen". Wir werden diese immer zusammen wählen. Die **Grundidee** hiervon und von den weiteren Punkten ist: Man untersucht denjenigen Ausdruck, der am sinnvollsten erscheint unter der Hypothese $\mathcal{H}_0$ und sucht dann *sinnvoll* den Ablehnungsbereich.

9.2 t-Test

Storrer-Bsp 45.2.A zur Qualitätskontrolle: Brote mit Gewichten [g]

$$69, 70, 71, 68, 67, 70, 70, 70, 67, 69$$

Bäcker: "Brote wiegen im Schnitt 70 g." Der Durchschnitt von obigen Broten ist jedoch 69.1 g. Wir einigen uns darauf, das Gewicht der Brote mit einer Normalverteilung zu modellieren. 2 Bilder zur Grundproblematik:

9.2.1 σ^2 vorerst bekannt, unrealistisch

Angenommen, wir haben eine Stichprobe $x_1, \ldots, x_n$, von der wir wissen, dass sie aus einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ stammt. Wir kennen leider nicht den Erwartungswert. Jemand behauptet, der Erwartungswert sei 70 ($\mathcal{H}_0 : \mu = 70, \mathcal{H}_1 : \mu \neq 70$, zweiseitig). Wie könnten wir das testen?

1. Hypothesen aufstellen:

2. α und (wenn möglich) n wählen:

3. gute Teststatistik wählen, 4. Verteilung der Teststatistik unter $\mathcal{H}_0$?

5. Kritischer Wert a

9.2.2 σ^2 unbekannt $\Rightarrow$ t-Test

Angenommen, wir haben eine Stichprobe $x_1, \ldots, x_n$, von der wir wissen, dass sie aus einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ stammt. Wir kennen leider weder Erwartungswert noch Varianz. Jemand hat behauptet, der Erwartungswert sei 70 ($\mathcal{H}_0 : \mu = 70, \mathcal{H}_1 : \mu \neq 70$, zweiseitig). Wie könnten wir das testen?

1. Hypothesen aufstellen:

2. α und (wenn möglich) n wählen:

3. gute Teststatistik wählen (vgl 9.2.1), 4. Verteilung der Teststatistik unter $\mathcal{H}_0$?

5. Kritischer Wert a

Storrer Bsp 45.2.A fertig besprechen.

Aufgaben aus Storrer (45.∞).

```
> b<-c(-2.2, -0.7, 3, 1.5, 0.2)
> t.test(b)
One Sample t-test
data:  b
t = 0.4028, df = 4, p-value = 0.7077
alternative hypothesis:  true mean is not equal to 0
95 percent confidence interval:
-2.121154 2.841154
sample estimates:
mean of x
0.36
```

9.2.3 2-Stichproben t-Test

Alte Sorte	410	420	430	440	450	450	480
Neue Sorte	440	450	455	480	490	505	

Angenommen, wir haben eine Stichprobe $x_1, \ldots, x_m$, von der wir wissen, dass sie aus einer Normalverteilung $\mathcal{N}(\mu_1, \sigma^2)$ stammt. Zudem haben wir eine Stichprobe $y_1, \ldots, y_n$, von der wir wissen, dass sie aus einer Normalverteilung $\mathcal{N}(\mu_2, \sigma^2)$ stammt. Die beiden Stichproben seien unabhängig voneinander, die Varianzen seien gleich - aber unbekannt! Wir wollen mit einem Test die beiden Mittelwerte vergleichen.

1. Hypothesen aufstellen:

2. α und (wenn möglich) m, n wählen:

3. gute Teststatistik wählen (vgl 9.2.2), 4. Verteilung der Teststatistik unter $\mathcal{H}_0$?

5. Kritischer Wert a

Falls Hinweis, dass $\sigma_1 \neq \sigma_2$, macht man den sogenannten Welch-Test.

Beispiel aus Storrer 46.2.A.

9.3 (vgl 8.3) Die 4 wichtigsten Tests aus der Praxis

9.4 χ^2-Tests (Anpassung und Unabhängigkeit)

Mehr zu 9.4 in Stahel Kapitel 10 (empfohlen, sind nur 12 Seiten).

9.4.1 χ^2-Anpassungstest

Wir motivieren die Grundfrage mit einem historischen Beispiel. Aufgrund der Mendelschen Erbgesetze erwartet man in bestimmten Kreuzungsversuchen, dass Nachkommen in drei verschiedenen Typen mit Wahrscheinlichkeiten 0.25 (aa), 0.5 (aA) und 0.25 (AA) auftreten.

In einer Untersuchung (Daten aus Stahel) fand man: $s_1 = 29, s_2 = 77$ und $s_3 = 44$ bei 150 Nachkommen. Wäre es exakt "fair" gewesen, so "müsste" es sein: $150/4 = 37.5$, $150/2 = 75$ und wieder $150/4 = 37.5$. Offenbar gibt es im Versuch eine Abweichung gegenüber diesen "fairen" Werten und jedeR WissenschaflerIn, welche mit genau diesen "fairen" Werten aufwarten würde, muss als Statistik-Banause bezeichnet werden (selbst wenn er/sie bei den halben Fällen *je einmal* auf- und abrundet). *Eine gewisse Variation in den Daten ist zu erwarten! Aber wieviel ist noch mit dem Modell von Mendel vereinbar? Ab wann müssen wir von einer signifikanten Abweichung sprechen?*

Wir müssen dazu erstmal eine neue Zufallsgrösse einführen, welche zu Recht an die Binomialverteilung erinnert: die **Multinomialverteilung** (die Binomialverteilung ist ein Spezialfall hiervon). Wenn wir n unabhängige Versuche haben mit m möglichen Ereignissen (in obigem Beispiel: $n = 150, m = 3$ - bei der Binomialverteilung $m = 2$ (Erfolg/Misserfolg)), so ist ein Zufallsvektor $(S_1, \ldots, S_m)$ per Definitionem multinomialverteilt, wenn

$$P[S_1 = s_1, \ldots, S_m = s_m] = \frac{n!}{s_1! \ldots s_m!} p_1^{s_1} \ldots p_m^{s_m}.$$

Dabei muss dann gelten: $s_1 + \ldots + s_m = n$ und $p_1 + \ldots + p_m = 1$.

Die Hypothesen sind derart, dass der Zufallsvektor $(S_1, \ldots, S_m)$ unter $\mathcal{H}_0$ eine Multinomialverteilung mit Wahrscheinlichkeiten $p_1, \ldots, p_m$ hat gegen $\mathcal{H}_1$ einer Multinomialverteilung mit Wahrscheinlichkeiten $q_1, \ldots, q_m$, wo mindestens bei zwei (!) Wahrscheinlichkeitspaaren (p_i, q_i), $1 \leq i \leq m$, gelten muss $p_i \neq q_i$.

Wie bei der Binomialverteilung wird man auch hier die Parameter sinnvollerweise mit

$$\hat{p}_i := s_i/n$$

schätzen. Damit erhalten wir m Abweichungen

$$\hat{p}_i - p_i, \ 1 \leq i \leq m.$$

Positive und negative Abweichungen könnten sich aufheben. Somit ist es sinnvoll, z.B. die absoluten Werte zu betrachten:

$$|\hat{p}_i - p_i| = |s_i - np_i|/n, \ 1 \leq i \leq m.$$

Wir wollen diese m Werte alle zusammen in einer einzigen Testgrösse betrachten und es hat sich herausgestellt, dass

$$u := \sum_{i=1}^{m} \frac{(s_i - np_i)^2}{np_i} \qquad (\chi^2 - \text{Anpassung})$$

eine brauchbare Testgrösse ist. Mit $n \to \infty$ gilt, dass diese Grösse als Zufallsgrösse U unter $\mathcal{H}_0$ eine χ^2_{m-1}-Verteilung besitzt. In Stahel sind Praktikerregeln aufgeführt, ab wann wir n als genügend gross bezeichnen dürfen. Es ist noch beachtenswert, dass in der Referenzverteilung nur m und nicht die einzelnen p_i's vorkommen. Wann werden wir die $\mathcal{H}_0$-Hypothese ablehnen, das heisst, von welcher Form ist der Ablehnungsbereich? Wo war wohl Mendel?

Wir untersuchen jetzt mit diesem Test das einführende Beispiel:

$$\begin{aligned} u &= \frac{(29 - 150 * 0.25)^2}{150 * 0.25} + \frac{(77 - 150 * 0.5)^2}{150 * 0.5} \\ &\quad + \frac{(44 - 150 * 0.25)^2}{150 * 0.25} \\ &= 1.927 + 0.053 + 1.127 \\ &= 3.11. \end{aligned}$$

Es gilt: $P[\chi^2_2 > 5.99] = 0.05$; somit werden wir auf dem $\alpha = 5\%$-Niveau die $\mathcal{H}_0$-Hypothese nicht ablehnen. Die Versuchsresultate sind demnach auf diesem Niveau durchaus mit den Mendel'schen Gesetzen vereinbar.

Im Storrer wird noch der Fall behandelt, dass die Werte der Parameter (hier 0.25, 0.5 und 0.25) nicht bekannt sind und aus den Daten geschätzt werden müssen. Zum Beispiel bei einer Po(λ)-Zufallsgrösse muss man noch λ schätzen. Dieser Stoff ist obligatorisch und kam schon in der Prüfung vor; bitte lesen.

9.4.2 χ^2-Test für Unabhängigkeit in Kontingenztafeln

Die folgenden Daten sind aus Radelet, M. (1981): "Racial Characteristics and the Imposition of the Death Penalty." Amer. Sociol. Rev. **46**: 918-927. 326 Personen sind alle des Mordes überführt worden - es ging noch darum, ob sie die Todesstrafe erhielten oder nicht. Die Daten betreffen 20 Counties in Florida von 1976-1977.

Total sind es 326 Personen. Von den 166 schwarzen Angeklagten wurden 17 zum Tode verurteilt. Von den 160 weissen Angeklagten wurden 19 zum Tode verurteilt. Gibt es einen statistischen Hinweis darauf, dass die Hautfarbe einen Einfluss auf das Urteil gehabt hat?

Obige Frage muss mit diesen (groben) Daten klar mit "Nein" beantwortet werden. Wenn es überhaupt eine signifikante Benachteiligung gäbe, dann wäre sie eher zu Lasten der Weissen als der Schwarzen. Es ist jedoch überraschenderweise doch so, dass die Schwarzen eindeutig benachteiligt sind! Wie ist so etwas möglich?

Wir motivieren die Grundfrage mit der (zu einfachen) Analyse der Todesurteile in den USA; im Storrer sind weitere Beispiele vorhanden. Am Besten trägt man die Daten in einer sogenannten Vierfeldertafel ein (und zwar die absoluten Werte!; konkret und abstrakt):

Wir haben den Merkmalen gleich Bezeichnungen gegeben, welche an Zufallsgrössen erinnern sollten: X und Y. Wir stellen jetzt ein Wahrscheinlichkeitsmodell auf:

$$P[X = i, Y = j] =: p_{ij}.$$

Die obigen Anzahlen n_{ij} in den einzelnen Kästchen fassen wir auch als Realisationen von Zufallsgrössen N_{ij} auf $(1 \leq i, j \leq 2)$. Der Zufallsvektor $(N_{11}, \ldots, N_{22})$ ist multinomialverteilt:

$$P[N_{11} = n_{11}, .., N_{22} = n_{22}] = \frac{n!}{n_{11}!..n_{22}!} p_{11}^{n_{11}} ..p_{22}^{n_{22}}.$$

Dabei muss dann gelten: $n_{11} + \ldots + n_{22} = n$ und $p_{11} + \ldots + p_{22} = 1$. Damit sind wir also in der gleichen Situation wie beim χ^2-Anpassungstest.

Was uns hingegen jetzt interessiert, ist, ob die beiden Ausprägungen unabhängig voneinander auftreten (im Beispiel: Es besteht kein Zusammenhang zwischen Hautfarbe und Wahrscheinlichkeit für ein Todesurteil). Formalisiert wird das in der Forderung (Nullhypothese)

$$p_{ij} := P[X = i, Y = j] = P[X = i]P[Y = j] =: p_{i.}p_{.j}.$$

Von (χ^2-Anpassung) erhalten wir jetzt folgende Teststatistik

$$u^{(\text{prov})} := \sum_{i,j} \frac{(N_{ij} - np_{i.}p_{.j})^2}{np_{i.}p_{.j}}. \qquad (\chi^2 - \text{Test auf Unabhängigkeit, provisorisch})$$

Die $p_{i.}$ und $p_{.j}$ sind unbekannt und wir müssen sie aus den Daten mit $\hat{p}_{i.} := N_{i.}/n$, $\hat{p}_{.j} := N_{.j}/n$ schätzen. Dadurch wird obiger Ausdruck zur definitiven Variante:

$$u := \sum_{i,j} \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n}. \qquad (\chi^2 - \text{Test auf Unabhängigkeit})$$

Man kann zeigen, dass diese Grösse als Zufallsgrösse U und bei Unabhängigkeit der Merkmale (Nullhypothese) mit $n \to \infty$ eine χ^2-Verteilung mit 1 Freiheitsgrad besitzt.

Im einleitenden Beispiel erhalten wir einen Wert $u = 0.2214$. Der P-Wert von 0.2214 bei einer χ^2-Verteilung mit 1 Freiheitsgrad ist 0.638 (hab ich vom Computer); bei $\alpha = 5\%$ erhalten wir einen Ablehnungsbereich der Nullhypothese ab Werten $u \geq 3.841$. Was schliessen wir daraus? Was sollten wir nicht schliessen?

Warum *Freiheits*grad?

9.5 Schlussbemerkungen zu Tests

Vergleich Tests und KI's: Sie haben mittlerweile, vor allem beim t-test, gemerkt, dass bei Tests und KI's die gleichen Zahlen vorkommen. Dazu drei Bemerkungen:

* Es gibt operationell einen Unterschied: bei den KI's sind keine Hypothesen involviert; wir arbeiten nur mit den Daten und vergleichen keineswegs mit allenfalls vorhandenen Theorien und ihren Parametern (KI's kamen ja auch schon in Kapitel 8 und die Hypothesen erst in Kapitel 9).
* Da meist die beiden Berechnungen als Teilaufgaben in einer Aufgabe vorkommen, lohnt es sich, dies zu merken, damit man nicht alles zweimal berechnet - auch in R wird meist automatisch beides angegeben.
* Das alles ist kein Zufall: Tests sind umgestülpte KI's, mehr dazu in (45.6).

Eine kleine Betrachtung

Wichtig:

1. Lesen Sie jetzt das komplette Kapitel im Storrer II selber durch (Kapitel 44-47).
2. Lösen Sie danach mindestens 5 Aufgaben hinten im Kapitel und vergleichen Sie mit den Lösungen am Schluss des Buches. Bei Bedarf lösen Sie mehr Aufgaben.
3. Gehen Sie in die Übungsstunde. Drucken Sie das Übungsblatt dazu *vorher* aus, lesen Sie *vorher* die Aufgaben durch und machen sich erste Gedanken dazu (zum Beispiel, wie man sie lösen könnte).
4. Dann lösen Sie das Übungsblatt: zuerst immer selber probieren, falls nicht geht: Tipp von Mitstudi benutzen, falls immer noch nicht geht: Lösung von Mitstudi anschauen, 1 Stunde warten, versuchen, aus dem Kopf heraus wieder zu lösen, falls immer noch nicht geht: Lösung von Mitstudi abschreiben (und verstehen - also sollte man insbesondere keine Fehler abschreiben!).
5. Lösen Sie die entsprechenden Prüfungsaufgaben im Archiv.