

Mathematik II für Biologen

Übungsblatt 7 (Abgabe am 12.06.2009)

Aufgabe 18 (χ^2 -Test, sprich: *chi-Quadrat-Test*) (10 Punkte)

In Großbritannien wurden 2588 verschiedene Gebiete von je 10 km² Größe untersucht⁴ und nach der Häufigkeit bestimmter Schmetterlinge als “Schmetterlings-Hotspot”, “Schmetterlings-Coldspot” oder “neutrales Gebiet” klassifiziert. Aufgrund theoretischer Überlegungen sollten 5% der Gebiete Hotspots sein, 5% Coldspots und die übrigen 90% neutral.

- a) Wieviel Gebiete von jeder Sorte sollte es somit geben, falls die tatsächlichen Zahlen so exakt wie möglich mit den theoretischen Vorhersagen übereinstimmen? (Dies sind die “erwarteten” Anzahlen.)

Nun kann man jedoch nicht erwarten, dass die theoretischen und die empirischen Zahlen genau miteinander übereinstimmen, so, wie auch nicht erwartet werden kann, dass bei 120 Würfeln eines fairen Würfels jede Augenzahl genau 20 Mal geworfen wird. Damit die Theorie als weiterhin vertretbar gilt, wird vielmehr verlangt, dass sich die Natur so verhält, als würde sie aus einem sehr, sehr großen Vorrat an Gebieten, von denen genau 5% Hotspots, genau 5% Coldspots und genau 90% neutral sind, ohne Zurücklegen 2588 Gebiete ziehen und auf Großbritannien verteilen. Falls sich die Natur so verhält, können die konkreten, gemessenen Anzahlen zwar um die in Teil a berechneten schwanken, jedoch sind “im Durchschnitt” alle 3 Gebietstypen in den richtigen Anteilen vertreten.

In der Praxis war es denn auch so, dass 123 Hotspots, 147 Coldspots und 2318 neutrale Gebiete gezählt wurden, was nicht genau den in Teil a berechneten Zahlen entspricht. (Dies sind die “beobachteten” Anzahlen.) Reichen diese Abweichungen aus, um auf dem Signifikanz-Niveau von $\alpha = 5\%$ die Nullhypothese H_0 , dass die Theorie stimmt, zugunsten der Alternativhypothese H_A , dass sie nicht stimmt, zu verwerfen? Dazu wählen wir als Teststatistik

$$\chi^2 := \sum_{i=1,2,3} \frac{(\text{beobachtete Anzahl Gebiete der Sorte } i - \text{erwartete Anzahl Gebiete der Sorte } i)^2}{\text{erwartete Anzahl Gebiete der Sorte } i}.$$

(Vgl. mit T aus Aufgabe 17.) Je größer der Wert von χ^2 desto größer der Verdacht, dass H_A gilt. Man könnte nun wie in Aufgabe 17 den p-Wert mittels Monte-Carlo-Simulation schätzen und dann je nachdem, ob der (geschätzte) p-Wert kleiner als α ist oder nicht, H_0 verwerfen oder nicht. Für den Fall $\alpha = 5\%$ kann man jedoch stattdessen folgende Faustregel⁵ für den Verwerfungsbereich anwenden:

Verwerfe H_0 auf dem Signifikanz-Niveau $\alpha = 5\%$, falls $\chi^2 > \nu + 2\sqrt{2\nu}$, wobei ν die sog. Anzahl der Freiheitsgrade ist.

In dieser Art von Problem ist die Anzahl ν der Freiheitsgrade gleich der Anzahl der betrachteten Klassen minus 1. Da es hier drei Klassen von Gebieten gibt, ist somit $\nu = 3 - 1 = 2$.

- b) Berechnen Sie diese “kritische Grenze” $\nu + 2\sqrt{2\nu}$.
c) Berechnen Sie den Wert der Teststatistik χ^2 .
d) Wie lautet demnach die Testentscheidung?
e) Wiederholen Sie c und d für den Fall, dass 120 Hotspots, 150 Coldspots und 2318 neutrale Gebiete gefunden worden wären.
f) Testen Sie mit obigem χ^2 -Test ohne Simulation, ob die Daten aus Aufgabe 17 auf dem Signifikanz-Niveau $\alpha = 5\%$ mit Mendels Hypothese vereinbar sind.

⁴siehe: W. Mendenhall & T. Sincich: *Statistics for Engineering and the Sciences*, 4th ed., Prentice Hall, Exercise 10.9 und J.R. Prendergast et al. *Rare species, the coincidence of diversity hotspots and conservation strategies* Nature **365** (1993) 335–337, doi:10.1038/365335a0

⁵Diese Faustregel kann angewandt werden, wenn für etwa 80% der betrachteten Klassen gilt, dass die erwartete Anzahl jeweils ≥ 4 ist, und für die restlichen Klassen die erwartete Anzahl jeweils ≥ 1 ist. Dies ist hier erfüllt.

Aufgabe 19

(10 Punkte)

Man kann die χ^2 -Teststatistik aus Aufgabe 18 auch benutzen, um zu testen, ob die Daten zu gut zum Modell passen, um echt zu sein.

- a) (MATLAB) Angenommen jemand behauptet, er hätte 900 Mal einen fairen Würfel geworfen und dabei jede Augenzahl genau 150 Mal beobachtet. Würden Sie ihm glauben? Schliesslich entsprechen die beobachteten Anzahlen n_i von Würfeln der Zahl i genau den für einen fairen Würfel erwarteten Zahlen, d.h. die Teststatistik χ^2 ist 0. Oder ist die Übereinstimmung hier zu gut, um wahr zu sein? Führen Sie das Experiment, einen fairen Würfel 900 mal zu werfen, n -mal durch (n groß) und überprüfen Sie, in wievielen Fällen $\chi^2 = 0$ beobachtet wurde. Interpretieren Sie das Ergebnis. (Hier ist sozusagen H_0 : *Daten sind echt* und H_A : *Daten sind gefälscht*.)

```
>> n=10000; % n kleiner waehlen, falls Rechner zu langsam
>> chi_quadrat=zeros(1,n);
>> anzahl=zeros(1,6);
>> for k=1:n
    wuerfe=unidrnd(6,1,900);
    for i=1:6
        anzahl(i)=sum(wuerfe==i);
    end
    chi_quadrat(k)=sum((anzahl-[150 150 150 150 150 150]).^2/150);
    % Ob man hier noch durch 150 teilt oder nicht, spielt keine Rolle.
end
>> sum(chi_quadrat==0)/n % sum(T==0) zaehlt, wie oft T==0 beobachtet wurde.
```

- b) (MATLAB) *Hat Mendel geschummelt?* Ermitteln Sie mittels Monte-Carlo-Simulation, wie groß in etwa die Wahrscheinlichkeit dafür ist, bei Gültigkeit der Mendelschen Gesetze für 556 Erbsen einen Wert von χ^2 zu erhalten, der *kleiner* ist als der in Aufgabe 17 b berechnete. Spricht dies dafür, dass Mendel seine Zahlen geschönt hat?
- c) Mendel hat viele Versuche durchgeführt, nicht nur den oben erwähnten. Angenommen, in Teil b konnte gezeigt werden, dass die Zahlen in obigem Beispiel so gut zu den theoretisch zu erwartenden passen, dass es statistisch signifikant auf dem Signifikanz-Niveau von $\alpha = 5\%$ ist, dass sie nicht mehr rein zufällig zustande gekommen sein können. Welches Argument könnte man dennoch zur Verteidigung von Mendel anführen?

Aufgabe 20

(10 Punkte)

Beim Spiel Würfelzwerge von Selecta gibt es Kärtchen, auf denen Zwerge mit drei Kleidungsstücken (Mütze, Jacke und Hose) dargestellt sind. Die Kleidungsstücke kommen in 6 unterschiedlichen Farben (rot, gelb, grün, blau, lila und pink) vor. Man würfelt mit drei Farbwürfeln und muss schnell einen Zwerg finden, der Kleidung in der entsprechenden Farbkombination trägt. Zu jeder möglichen Farbkombination gibt es genau einen Zwerg, d.h., wenn es einen Zwerg mit gelber Mütze, gelber Jacke und blauer Hose gibt, dann kann es keinen mit gelber Mütze, blauer Jacke und gelber Hose geben, aber sehr wohl noch einen mit gelber Mütze, blauer Jacke und blauer Hose.

- a) Wieviele Kärtchen hat das Spiel?
- b) Wieviele ein-, zwei- und dreifarbige Zwerge gibt es jeweils?
- c) Wie groß ist jeweils die Wahrscheinlichkeit, mit fairen Farbwürfeln die Kombination für einen ein-, zwei- oder dreifarbigem Zwerg zu würfeln?

Aufgabe 21

(10 Punkte)

Ungefähr 2% der Bevölkerung leiden an rheumatoider Arthritis. Etwa 70% davon haben das humane Leukozyten Antigen HLA-DR4, das sonst nur 25% der allgemeinen Bevölkerung haben. Angenommen, eine zufällig ausgewählte Person hat HLA-DR4. Mit welcher Wahrscheinlichkeit leidet diese Person unter rheumatoider Arthritis?

Bitte definieren Sie, ähnlich wie im Vorlesungsbeispiel *Diagnostischer Test* zum Satz von Bayes, klar die Ereignisse und Wahrscheinlichkeiten, die es zu betrachten gilt, und zeigen Sie alle Rechenschritte.