

Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun

Heinrich-Heine-Universität Düsseldorf

28. Oktober 2010

1 Zufallsworte

- Unabhängigkeit
- Aufbau von Worten
- DNA
- Shotgun Methode

Stochastische Unabhängigkeit

- Zwei Ereignisse A und B heißen (*stochastisch*) *unabhängig*, wenn

$$P(A \cap B) = P(A) \cdot P(B)$$

- In konkreten Experimenten wird die Unabhängigkeit durch den Versuchsaufbau gewährleistet

Beispiel für abhängige Ereignisse

Zweifacher Wurf eines fairen Würfels

$A = \text{"Erster Wurf eine 3"}$

$B = \text{"Augensumme 8"} = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$

$A \cap B = \{(3, 5)\}$

Folgende Wahrscheinlichkeiten

$$P(A) = \frac{1}{6} \quad P(B) = \frac{5}{36} \quad P(A \cap B) = \frac{1}{36}$$

Aber

$$P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{5}{36} = 0.02315 \neq 0.02778 = \frac{1}{36}$$

Also sind die beiden Ereignisse stochastisch abhängig.

Erfolglose unabhängige Versuchswiederholungen

- Derselbe Versuch wird unabhängig n -mal wiederholt
- Jeder einzelne Versuch gelingt mit Wahrscheinlichkeit p
- Mit welcher Wahrscheinlichkeit misslingen alle n Versuche?
- Ein einzelner Versuch misslingt mit Wahrscheinlichkeit $1 - p$
- Zwei unabhängige Versuche misslingen mit Wahrscheinlichkeit $(1 - p)^2$ gemeinsam
- n unabhängige Versuche misslingen mit Wahrscheinlichkeit $(1 - p)^n$ gemeinsam

Beispiel: Mindestens ein Erfolg

- 500 000 Versuche mit Erfolgswahrscheinlichkeit $p = 0.00002217$
- Mit welcher Wahrscheinlichkeit gibt es mindestens einen Erfolg?
- Übergang zum Komplement: Mit welcher Wahrscheinlichkeit gibt es keinen Erfolg?
- Misserfolgswahrscheinlichkeit im Einzelfall $1 - p = 0.9999778$
- Wahrscheinlichkeit von 500 000 unabhängigen Misserfolgen:

$$(1 - p)^{500\,000} = 0.9999778^{500\,000} = 0.000015$$

- Wahrscheinlichkeit mindestens eines Erfolges

$$1 - 0.000015 = 0.999985$$

Zufallsworte

- Gegeben ist ein Alphabet
- Jeder Buchstabe des Alphabets besitzt eine Wahrscheinlichkeit
- Zufallsworte entstehen dadurch, dass Buchstaben unabhängig gezogen und aneinander gereiht werden
- Die Wahrscheinlichkeit eines Worts ist gleich dem Produkt der Wahrscheinlichkeiten seiner Buchstaben

Alphabet m, w

- Alle Worte der Länge vier aus dem Alphabet $\{m, w\}$

$$\Omega = \{(m, m, m, m), (m, m, m, w), (m, m, w, m), (m, m, w, w), \\ (m, w, m, m), (m, w, m, w), (m, w, w, m), (m, w, w, w), \\ (w, m, m, m), (w, m, m, w), (w, m, w, m), (w, m, w, w), \\ (w, w, m, m), (w, w, m, w), (w, w, w, m), (w, w, w, w)\}$$

- Modelliert man das Geschlechterverhältnis wie in der Mathematik, dann tragen m und w beide die Wahrscheinlichkeit 0.5
- Alle Worte haben dann dieselbe Wahrscheinlichkeit, nämlich $\frac{1}{16} = 0.0625$

Alphabet m, w, Fortsetzung

- In der Biologie hat m die Wahrscheinlichkeit 0.4 und w hat 0.6
- Alle Worte mit 2-mal m und 2-mal w haben die Wahrscheinlichkeit $0.4^2 \cdot 0.6^2 = 0.0576$
- Also ist die Wahrscheinlichkeit, dass auf zufällig ausgewählten 4 Sitzen in dieser Vorlesung 2 männliche und 2 weibliche Studierende sitzen, gleich $6 \cdot 0.0576 = 0.3456$

Das genetische Alphabet

- Vier verschiedene Buchstaben mit unterschiedlicher Häufigkeit

Base	Häufigkeit
a	40%
c	22%
g	17%
t	21%

- Beispiele:
 - $P(acac) = 0.4 \cdot 0.22 \cdot 0.4 \cdot 0.22 = 0.007744$
 - $P(gtgt) = 0.17 \cdot 0.21 \cdot 0.17 \cdot 0.21 = 0.001275$

cg

- Untersuche Kombinationen zweier aufeinander folgender Basen
- In Organismen tritt die Kombinationen *cg* nur in 1% aller Fälle auf
- In einem Zufallswort wäre die Häufigkeit

$$0.22 \cdot 0.17 = 0.0374$$

- Bei den anderen Dinukleotiden ist der Unterschied weniger ausgeprägt.

DinoDNA aus Crichton's Jurassic Park

gcgttgctggcggtttttccataggctccgccccctgacgagcatcacaaaaatcgacgc
ggtggcgaaacccgacaggactataaagataaccaggcggtttccccctggaagctccctcg
tgttccgaccctgccgcttaccggatacctgtccgcctttctcccttcgggaagcgtggc
tgctcacgctgtacctatctcagttcgggtgtaggtcgttcgctccaagctgggctgtgtg
ccgttcagcccgaccgctgcgcccttatccggtaactatcgtcttgagtccaaccggtaa
agtaggacaggtgccggcagcgctctgggtcattttcggcgaggaccgctttcgctggag
atcggcctgtcgcttgcggtattcggaatcttgacgcccctcgctcaagccttcgtcact
ccaaacgtttcggcgagaagcaggccattatcgccggcatggcggccgacgcgctgggct
ggcgttcgcgacgcgaggctggatggccttcccattatgattcttctcgcttcggcg
cccgcgttgacaggccatgctgtccaggcaggtagatgacgaccatcagggaacagcttcaa
cggctcttaccagcctaacttcgatcactggaccgctgatcgtcacggcgatttatgccg
cacatggacgcgttgctggcggtttttccataggctccgccccctgacgagcatcacaaa
caagttagaggtggcgaaacccgacaggactataaagataaccaggcggtttccccctggaa
gcgctctcctgttccgaccctgccgcttaccggatacctgtccgcctttctcccttcggg
ctttctcaatgctcacgctgtaggtatctcagttcgggtgtaggtcgttcgctccaagctg
acgaacccccggttcagcccgaccgctgcgcccttatccggtaactatcgtcttgagtcca
acacgacttaacgggttggcatggattgtaggcgccgccctataccttgctgcctcccc
gcggtgcatggagccggggccacctcgacctgaatggaagccggcgccacctcgctaacgg
ccaagaattggagccaatcaattcttgcgggagaactgtgaatgcgcaaaccaacccttgg
ccatcgcgtccgccatctccagcagccgcacgcggcgcatctcgggcagcggttgggtcct

[illegible]

Auswertung der DinoDNA

- Anteil c: 33%
- Anteil g: 27%
- Anteil cg: 9.6%
- $0.33 \cdot 0.27 = 0.089$

Die Sequenz aus dem Buch von Crichton ist vermutlich ein Zufallswort.

Shotgun Methode

- DNA soll sequenziert werden
- Die zu untersuchende Sequenz wird vervielfältigt und dann zerstückelt.
- Am Ende liegen beispielsweise 500 000 Schnipsel vor.
- Ein Schnipsel endet mit ttaagatc, ein Schnipsel beginnt mit ttaagatc. Mit welcher Wahrscheinlichkeit setzt der zweite Schnipsel den ersten fort?

Schnipselbeispiel, durchgerechnet

- Modellannahme für diese Rechnung: Zufallswort der Länge 8

$$P(\text{ttaagatc})$$

$$= 0.21 \cdot 0.21 \cdot 0.40 \cdot 0.40 \cdot 0.17 \cdot 0.40 \cdot 0.21 \cdot 0.22$$

$$= 0.00002217$$

- In der Probe befinden sich ein paar Dutzend Schnipsel, die entweder Kopien der untersuchten Sequenz sind oder sie fortsetzen
- Jede andere Übereinstimmung ist zufällig
- Oben ausgerechnet: Die Wahrscheinlichkeit einer zufälligen Übereinstimmung beträgt 0.999985

Shotgun Methode, Fortsetzung

- Wie lang müssen die Endstücke sein, damit aus ihrer Übereinstimmung darauf geschlossen werden kann, dass das eine Stück die Fortsetzung des anderen ist?
- Die Wahrscheinlichkeit, dass zwei Stücke zufällig dieselben n Basen am Ende haben, soll kleiner sein als $0.0001/500\,000$
- Am schwierigsten für aaa ... aa (n -mal)

$$P(\underbrace{\text{aaa} \dots \text{aa}}_{n\text{-mal}}) = 0.40^n$$

Shotgun Methode, Fortsetzung

Gesucht ist n mit

$$0.40^n \leq 2 \cdot 10^{-10}$$

Löse durch Logarithmieren:

$$n \cdot \ln(0.40) = \ln(2) - 10 \cdot \ln(10)$$

$$n = \frac{\ln(2) - 10 \cdot \ln(10)}{\ln(0.40)} = \frac{-22.33}{-0.9163} = 24.37$$

Man verlangt, dass 25 Basen übereinstimmen.