ÜBUNGEN ZUR WAHRSCHEINLICHKEITSRECHNUNG UND STATISTIK FÜR BIOLOGEN

Blatt 8

---

1. Aufgabe Die Bohnenspinnmilbe ist ein weitverbreiteter Schädling bei Nutzpflanzen, auf den die Pflanzen durch die Produktion toxischer Substanzen reagieren. Um zu prüfen, ob sich Baumwollpflanzen an Schädlingsbefall „erinnern“, wurden zwei Gruppen von je 20 Baumwollpflanzen mit Milben infiziert: Die erste Gruppe war noch nie von Bohnenspinnmilben befallen worden, die Pflanzen der zweiten Gruppe hatten bereits Milbenbefall er- (und über-)lebt. Nach einiger Zeit wurde die Anzahl Milben auf den Pflanzen ausgezählt, die Ergebnisse finden Sie in der Datei `milben.csv` (simulierte Daten).
Sind die Befallszahlen in den beiden Gruppen signifikant unterschiedlich? Stellen Sie die Ergebnisse graphisch dar, führen Sie einen passenden t-Test durch und formulieren Sie eine Antwort aus.

2. Aufgabe Es wird vermutet, dass bei einem Pferderennen auf einem ovalen Parcours die Startposition einen Einfluß auf die Gewinnchance hat. In 144 Rennen hatten die Sieger die Startpositionen $1, 2, \ldots, 8$ mit den folgenden Häufigkeiten: $29, 19, 18, 25, 17, 10, 15, 11$. Testen Sie, ob die Startposition einen Einfluß auf die Siegchance hat, zum Irrtumsniveau 5%.

3. Aufgabe Ein HIV-Test habe eine Sensitivität von 99,9% (d.h. er schlägt bei 99,9% der Infizierten an) und eine Spezifizität von 95% (d.h. er schlägt bei 5% der nicht-Infizierten an). 8% einer Population seien infiziert. Wenn eine zufällig aus der Population gegriffene Person ein positives Testergebnis hat, wie wahrscheinlich ist es dann, dass sie tatsächlich infiziert ist?

4. Aufgabe (simulierte Daten, freundlicherweise von Prof. S. Diehl zur Verfügung gestellt)
Es könnte sein, dass Dickhorn-Mutterschafe mehr (oder weniger) Nahrung benötigen, wenn sie Lämmer säugen. Um dies zu untersuchen, wurde bei 16 weiblichen Schafen jeweils die mittlere Zeit (in Minuten pro Tag) bestimmt, die sie mit Grasen zubringen, einerseits während der Zeit des Säugens, andererseits in einem Referenzzeitraum ohne Lämmer. Es gab folgendes Ergebnis:

Schaf Nr.	12	17	23	24	29	31	33	48	55	57	59	60	63	71	73	78
Zeit mit Lamm	272	262	256	260	252	261	253	252	256	260	249	243	254	270	247	250
Zeit ohne Lamm	256	244	267	268	258	262	245	257	273	277	264	254	256	291	264	271

Grasen Mutterschafe mit Lämmern signifikant anders als ohne? Geben Sie auch ein Konfidenzintervall (beispielsweise zum Niveau 5%) für die Differenz der mittleren Futterzeiten an und formulieren Sie einen Antwortsatz.

5. Aufgabe Sie besuchen eine fremde Stadt, von der Sie wissen, dass die Taxis dort fortlaufend mit $1, 2, \ldots, N$ nummeriert sind, Sie kennen aber nicht die Gesamtzahl N. Während Sie vor dem Bahnhof auf den Bus warten, sehen Sie 12 Taxis vorbeifahren. Sie haben die Nummern 455, 190, 39, 542, 370, 289, 274, 237, 374, 899, 741 und 706.

(a) Wie groß ist die Wahrscheinlichkeit, genau diese Taxinummern in dieser Reihenfolge zu sehen, wenn $N = 912$ ist? (Modellieren Sie das Erblicken eines Taxis durch “Ziehen mit Zurücklegen”, d.h. ein Taxi kann prinzipiell auch zweimal gesehen werden.)

(b) Wie groß ist die Wahrscheinlichkeit, genau diese Taxinummern in dieser Reihenfolge zu sehen für beliebiges N? (Insbesondere: Was passiert mit der Wahrscheinlichkeit der Beobachtung für $N < 899$?)

(c) Berechnen Sie auf der Basis Ihrer Beobachtung den Maximum-Likelihood-Schätzer für N.

(d) Wie sähe der ML-Schätzer im allgemeinen Fall aus: Sie sehen n Taxis mit Nummern $x_1, x_2, \ldots, x_n$?

6. Aufgabe Auf einem Abschnitt nicht-kodierender DNS der Länge 1000 Basenpaare werden zwischen Mensch und Schimpanse $N = 23$ Mutationen gezählt. Ein Wissenschaftler argumentiert: „Die Anzahl Mutationen auf einem Stück der Länge 1000 ist Binomial-verteilt mit Erwartungswert $np \approx n\widehat{p} = 23$ und Standardabweichung $\sqrt{np(1-p)} \approx \sqrt{n\widehat{p}(1-\widehat{p})} \approx \sqrt{n\widehat{p}} = \sqrt{23}$. Mit der asympotischen Normalität folgt, dass $[13.6, 32.4]$ ein 95%-Konfidenzintervall für die erwartete Anzahl Unterschiede auf einem Stück DNS der Länge 1000 aus derselben genomischen Region ist."

a) Erklären Sie, wie dieses Intervall berechnet wurde.

b) Erzeugen Sie unter der Annahme, dass das wahre $p = 0.023$ ist, zufällige Anzahlen von Mutationen und berechnen Sie aus den simulierten Daten Konfidenzintervalle nach dieser Methode (Hinweis: der R-Befehl `rbinom` simuliert binomialverteilte Zufallsgrößen). Mit welcher Wahrscheinlichkeit enthält das so konstruierte Konfidenzintervall den wahren Wert?

c) Führen Sie Teil b) auch für $p = 0.007$ durch.