Universität Tübingen
Mathematisches Institut
Dr. Stefan Keppeler

Sommersemester 08

Mathematik II für Biologen

Übungsblatt 12 (Abgabe am 09.07.2008)

Aufgabe 40 (Fortsetzung von Aufgabe 37) (20 Zusatzpunkte)

→ Bitte bearbeiten Sie auch diese Aufgabe ohne MATLAB o.ä., Taschenrechner ist o.k. ←

Der Fettgehalt der Milch von neun verschiedenen Kühen derselben Rasse betrug, wie bereits berichtet, in Prozent

4,59 4,48 4,21 4,24 4,45 4,35 4,08 4,03 4,63.

a) Testen Sie mit einem **t-Test** auf dem 5%-Niveau die Nullhypothese, dass der theoretische, "wahre" mittlere Fettgehalt $\mu = 4.22\%$ ist gegenüber der Alternative
   (i) $\mu \neq 4{,}22\%$
   (ii) $\mu > 4{,}22\%$.
   Geben Sie hierbei die konkrete Formel für die Teststatistik an, deren Verteilung unter Annahme der Nullhypothese, das jeweilige Verwerfungskriterium, den beobachteten Wert der Teststatistik und die jeweilige Testentscheidung. Geben Sie außerdem jeweils das zugehörige 95%-Vertrauensintervall für μ an.

b) Testen Sie mit einem **Wilcoxon-Test** auf dem 5%-Niveau die Nullhypothese, dass der theoretische, "wahre" mittlere Fettgehalt $\mu = 4{,}22\%$ ist gegenüber der Alternative
   (i) $\mu \neq 4{,}22\%$
   (ii) $\mu > 4{,}22\%$.
   Berechnen Sie dazu die Teststatistik (Rechenweg zeigen!), bestimmen Sie das jeweilige Verwerfungskriterium und formulieren Sie die jeweilige Testentscheidung.

c) Mit welcher Wahrscheinlichkeit kann man mit dem **z-Test** aus Aufgabe 37 (ii) und $n = 9$ Daten statistisch auf dem Signifikanz-Niveau $\alpha = 5\%$ zeigen, dass $\mu > 4{,}22\%$ ist, wenn in Wirklichkeit $\mu = 4{,}4\%$ ist? (Dies ist die Macht des Tests, $1-\beta$, für die Nullhypothese $H_0 : \mu = 4{,}22$ gegenüber der Alternative $H_A : \mu = 4{,}4\%$.)

   HINWEIS: Erinnern Sie sich daran, dass in Aufgabe 37 (ii) H_0 genau dann verworfen wird, wenn

$$Z := \frac{S_n - 4{,}22\% n}{\sigma\sqrt{n}} > 1{,}64, \quad \text{wobei} \quad S_n := X_1 + \ldots + X_n \quad \text{ist.}$$

   Lösen Sie diese Ungleichung nach S_n auf. Wie ist, andererseits, S_n verteilt, wenn H_A gilt?

Aufgabe 41 (10 Punkte)

Die bereits aus den Aufgaben 35 und 39 bekannte Datei `fishy.dat` enthält in der ersten Spalte die Länge (in cm), in der zweiten Spalte das Gewicht (Einheit leider unbekannt) und in der dritten Spalte den DDT-Gehalt (in ppm) von $n = 96$ Welsen, die im Tennessee River in Alabama, USA, gefangen wurden. Bearbeiten Sie das Problem aus Aufgabe 39 (a) mit Hilfe eines zweiseitigen

a) t-Tests,
b) Wilcoxon-Tests,

d.h. testen Sie mit diesen Tests, ob die Daten mit der Annahme verträglich sind, dass das durchschnittliche Gewicht μ eines Welses aus diesem Fluss gleich 920 ist. MATLAB-Code:

```
>> load fishy.dat
>> gewicht=fishy(:,2)
>> [h,p,vi]=ttest(gewicht,920)
>> [p,h]=signrank(gewicht,920) % = Wilcoxon-Test
```

Interpretieren Sie die Ergebnisse.

Aufgabe 42 (Bootstrap, wird hier nochmal erklärt) (10 Zusatzpunkte)
Die Datei `m1bggws0708.dat` enthält in der ersten Spalte den Prozentsatz der positiv bewerteten Übungsaufgaben von 301 Studierenden der *Mathematik I für Biologen, Geowissenschaftler und Geoökologen* des Wintersemesters 07/08 und in der zweiten Spalte die zugehörige Klausurpunktzahl.

a) Zeichnen Sie ein Streudiagramm der Daten mit den Übungsprozenten auf der horizontalen Achse und den Klausurpunkten auf der vertikalen Achse. Tragen Sie auch die Regressionsgerade ein und berechnen Sie den (empirischen) Korrelationskoeffizienten (nach Pearson).
HINWEIS zu MATLAB: In das bereits gezeichnete Streudiagramm kann mit `lsline` die Regressionsgerade (nach der Methode der kleinsten Quadrate = least-squares-line) eingefügt werden.

Regressionsgerade und Korrelationskoeffizient scheinen anzudeuten, dass eine hohe Punktzahl in den Übungen tendenziell auch mit einer hohen Klausurpunktzahl einhergeht ($=H_A$). Wie sicher kann man sich dessen sein? Wir haben keinen Mechanismus, mit dem man weitere derartige Stichproben simulieren könnte, um dann zu sehen, wie oft die Steigung der Regressionsgeraden und der Korrelationskoeffizient negativ sind, denn wir wissen nicht, wie wir dem Computer beibringen sollen, die Übungs- und Klausurergebnisse eines Studierenden zu simulieren. Daher können wir auch nicht die Verteilung der Korrelation der Übungs- und Klausurergebnisse Ergebnisse von 301 Studierenden bestimmen, wenn die Nullhypothese H_0 gilt, dass die beiden Punktzahlen im wesentlichen nichts miteinander zu tun haben ("unkorreliert" sind), und können daher auch nicht entscheiden, ob der beobachtete Wert der Teststatistik (der Korrelationskoeffizient) groß genug ist, um H_0 zu verwerfen.

Um sich jedoch einen Eindruck davon zu verschaffen, wie stark die Korrelation für andere Gruppen von Studierenden hätte schwanken können, führen wir einen Bootstrap durch: Wir ziehen aus der ursprünglichen Stichprobe von 301 Datenpaaren mit Zurücklegen 301 Paare und berechnen davon die Teststatistik, die uns interessiert, also in diesem Fall die Korrelation. (Man beachte, dass die neue Stichprobe i.a. einige Zahlen, die in der ursprünglichen Stichprobe einfach auftauchten, nun mehrfach enthält und andere gar nicht.) Dies wiederholen wir n-mal (n groß) und hoffen dann, dass die Verteilung der Teststatistik, die man auf diese Weise erhält, so ähnlich ist wie die Verteilung, die man bekommen hätte, wenn man n Gruppen mit je 301 Studierenden die *Mathematik I* hätte durchlaufen lassen.

b) Zeichnen Sie ein Histogramm der so erhaltenen Korrelationskoeffizienten.

c) Was war der kleinste Korrelationskoeffizient, den Sie bei Ihrer Simulation beobachtet haben?

d) In etwa wievielen Fällen erhalten Sie einen positiven Korrelationskoeffizienten?

e) Könnte es Ihrer Meinung nach demnach dennoch so sein, dass der Korrelationskoeffizient für die Originaldaten nur eher zufällig positiv ist? HINWEIS: Bestimmen Sie aus Ihren Daten z.B. ein geeignetes Vertrauensintervall für den Korrelationskoeffizienten.

MATLAB Code dazu (unvollständig):

```
>> n=10000;
>> b=bootstrp(n,'corrcoef',p,k); % Zieht n-mal mit
    % Zuruecklegen je 301 Paare von p und k und berechnet
    % jeweils die Korrelation der neuen Stichprobe.
    % help bootstrp
>> hist(b(:,2),30)
>> min(b(:,2))
```