

Mathematik II für Biologen

Übungsblatt 13 (Abgabe am 11.07.2014)

Aufgabe 50 (Bootstrap, wird hier nochmal erklärt) (10 Punkte)

Die Datei `m1bggws1314.dat` enthält in der ersten Spalte die Übungspunkte von 179 Studierenden der Lehrveranstaltung *Mathematik I für Biologen, Geowissenschaftler und Geoökologen* aus dem Wintersemester 13/14 und in der zweiten Spalte die zugehörige Klausurpunktzahl.

- a) Zeichnen Sie ein Streudiagramm der Daten mit den Übungspunkten 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 von Studierenden zu simulieren. Daher können wir auch nicht die Verteilung der Korrelation der Übungs- und Klausurergebnisse von 179 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 179 Datenpaaren mit Zurücklegen 179 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 179 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) Bestimmen Sie ein 95%-Vertrauensintervall für den Korrelationskoeffizienten.
Sollten wir nun H_0 zugunsten von H_A verwerfen?
e) Führen Sie die gleiche Analyse auch für die entsprechenden Daten aus dem Wintersemester 09/10 (Datei: `m1bggws0910.dat`) durch. Gibt es einen signifikanten Unterschied? Worin könnte der Grund dafür liegen?

HINWEIS: Werfen Sie einen Blick auf die alten Vorlesungshompages unter
<http://www.maphy.uni-tuebingen.de/members/stke>

Hilfreiche MATLAB-Code-Schnipsel finden Sie auf der Rückseite.

MATLAB-Code zu Aufgabe 50 (unvollständig):

```
>> rng('shuffle')
>> load m1bggws1314.dat
>> corrcoef(???)
>> uebungen=m1bggws1314(:,1);
>> klausur=m1bggws1314(???)';
>> plot(uebungen,???,'+')
>> lsline
>> n=10000;
>> b=bootstrp(n,'???'',uebungen,klausur);
    % Zieht n-mal mit Zuruecklegen je 179 Paare von uebungen und klausur und
    % berechnet jeweils die Korrelation der neuen Stichprobe.
    % --> help bootstrp
>> hist(b(:,2),50)
>> B=sort(b(:,2));
>> stairs(B,(1:n)/n)
    % Was ist das, und wie bestimmt man daraus das gesuchte VI?
    % Geht alternativ auch mithilfe von mean(b(:,2)) und var(b(:,2)).
```

Aufgabe 51

(10 Punkte)

Im Einführungsbeispiel aus den Vorlesungen 3 & 4 haben wir eine Rinderherde von 1000 Tieren betrachtet. Durch Untersuchen einer Stichprobe von 10 Tieren (3 davon waren krank) wollten wir rückschließen, wieviele Tiere der Herde krank sind. Dazu haben wir Herden mit unterschiedlich vielen kranken Tieren simuliert und ein einseitiges 95%-Vertrauensintervall für die Anzahl kranker Tiere bestimmt: $\{88, 89, \dots, 1000\}$. Wir können dies auch als Vertrauensintervall $[8,8\%, 100\%]$ für den Anteil kranker Tiere interpretieren.

- a) Bestimmen Sie zum Vergleich mittels Bootstrap ein einseitiges 95%-Vertrauensintervall für den Anteil kranker Tiere.

ANLEITUNG:

- (i) Schätzen Sie den Anteil kranker Tier der Herde durch den Anteil kranker Tiere der Stichprobe, d.h. 30%. Nur mit welcher Genauigkeit? Hier kommt der Bootstrap ins Spiel.
 - (ii) Definieren Sie sich einen Datenvektor, der die beobachtete Stichprobe repräsentiert, z.B. einen Vektor mit drei Einsen und 7 Nullen. Wie erhält man daraus die Anzahl und den Anteil kranker Tiere an der Stichprobe?
 - (iii) Erzeugen Sie nun mittels Bootstrap viele (z.B. 100 000) ähnliche Stichproben, und ermitteln Sie für jede die Anzahl (oder den Anteil) kranker Tiere. Visualisieren Sie das Ergebnis, z.B. durch ein Histogramm, und bestimmen das gesuchte Vertrauensintervall.
- b) Bestimmen Sie nun auch – ebenfalls mittels Bootstrap – ein zweiseitiges 90%-Vertrauensintervall für den Anteil kranker Tiere.
- c) Hätten wir die Ergebnisse der Aufgabenteile a und b auch ermitteln können, ohne die Bootstrap-Stichproben tatsächlich zu erzeugen? Wie?