Lösungsvorschläge für die Übungsaufgaben der Vorlesung “Statistik“

Thomas Rupp
thomas@7t7.de

13. November 2001

Aufgabe 9

a)

Wir haben eine zufällige Kette $X_1, X_2, \ldots, X_n$ von unabh. Münzwürfen (n gross genug). Die Frage sein erstmal, wie wahrscheinlich es ist, dass wenn sie mit Kopf beginnt, auf Adler endet und die beiden Sequenzen der Köpfe am Anfang und Adler am Ende größer/gleich als m ist. Sei k_1' die Anzahl der aufeinanderfolgenden Köpfe am Anfang (k_1 die der Köpfe ohne den Vorgegebenen) und k_2 die der Adler am Ende.
Ganz allgemein gilt:

$$\begin{aligned} P(k \geq m) &= (1 - P(k \leq m-1)) = 1 - \sum_{i=0}^{m-1} P(k = i) \\ &= 1 - \sum_{i=0}^{m-1} 2^{-(i+1)} = 1 - \frac{2^m - 1}{2^m} = \frac{1}{2^m} \end{aligned}$$

Dann ist

$$\begin{aligned} & P(k_1' + k_2 \geq m, X_n = A | X_1 = K) \\ = \; & P(k_1 + k_2 \geq m - 1, X_n = A) \\ = \; & \sum_{i=0} P(k_1 = i) P(k_2 \geq (m-1) - i) + \sum_{i=m-1}^{n-2} P(k_1 = i) P(k_2 \geq 1) + P(k_1 = n-1) \\ = \; & \sum_{i=0}^{m-2} \left(\frac{1}{2}\right)^{i+1} \left(\frac{1}{2}\right)^{m-1-i} + \sum_{i=m-1}^{n-2} \left(\frac{1}{2}\right)^{i+1} \frac{1}{2} + \frac{1}{2^{n-1}} \cdot \frac{1}{2} \\ = \; & \sum_{i=0}^{m-2} \left(\frac{1}{2}\right)^{m} + \sum_{i=m-1}^{n-2} \left(\frac{1}{2}\right)^{i+2} + \frac{1}{2^n} \\ = \; & \frac{m-1}{2^m} + \frac{1}{2^n} + \sum_{i=m}^{n-1} \left(\frac{1}{2}\right)^{i+1} \\ = \; & \frac{m-1}{2^m} + \frac{1}{2^n} + \sum_{i=0}^{n-1} \left(\frac{1}{2}\right)^{i+1} - \sum_{i=0}^{m-1} \left(\frac{1}{2}\right)^{i+1} \\ = \; & \frac{m-1}{2^m} + \frac{1}{2^n} + \left(1 - \frac{1}{2^n}\right) - \left(1 - \frac{1}{2^m}\right) \\ = \; & \frac{m-1}{2^m} + \frac{1}{2^n} - \frac{1}{2^n} + \frac{1}{2^m} = \frac{m}{2^m} \end{aligned}$$

Dadurch bekommen wir für $m = 7$ das gesuchte Ergebnis $\frac{7}{2^7} \approx 0.0547$, für $m = 11$ jedoch schon ≈ 0.00537.

b)

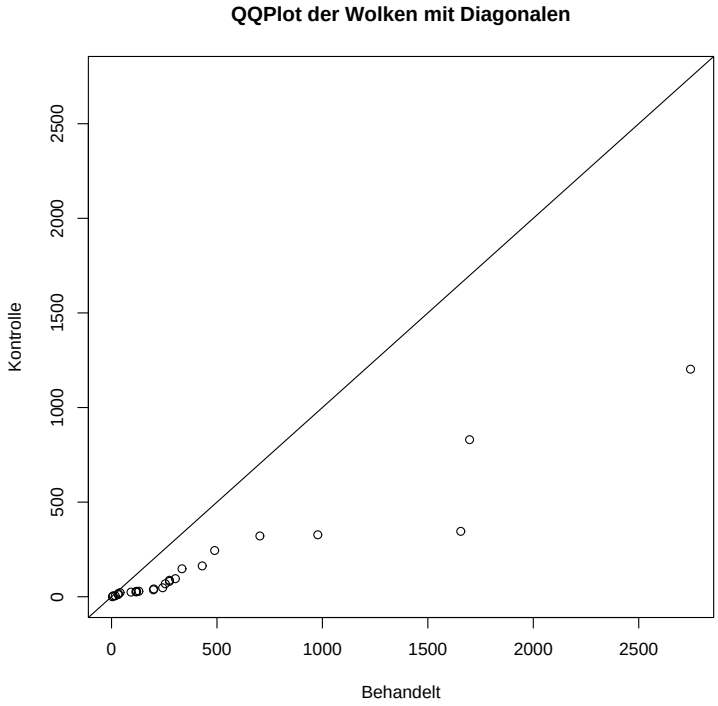
Interpretieren bzw. vergleichen können wir das Ganze etwa so: wir nehmen zwei gleich grosse Datensätze. Sie werden nach ihrer Herkunft mit "K" und "A" markiert. Jetzt werden beide in einen Topf geworfen und dann wird eine Stichprobe der Größe n gezogen.
Bei wirklich großen Datensätzen ist die Wahrscheinlichkeit einen bestimmten Typ zu ziehen in etwa 0.5 (ziehen ohne zurücklegen macht dann keinen Unterschied). Wir ordnen die Stichprobe der Größe nach; Tukeys Test schaut nun nach, wie wieviele der kleinsten Werte "hintereinander" aus dem einen Datensatz (K) kleiner sind[1],

[1] Als den ersten Datensatz K definieren wir den mit dem kleinsten Wert; daher ist das erste K fest.

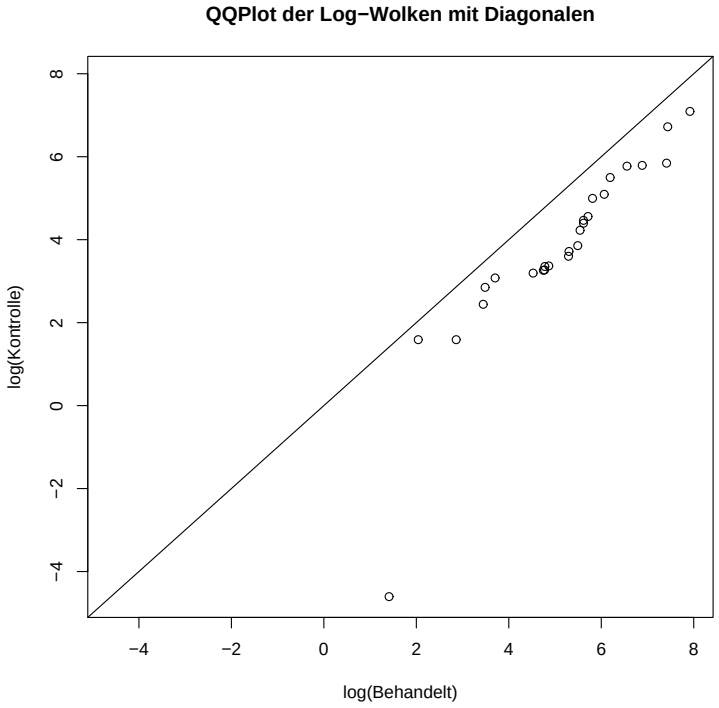
als der kleinste aus dem anderen Datensatz (A) gezogene. Dasselbe passiert mit den größten Werten am Stück aus dem anderen Datensatz (A), die mit dem größten aus dem einem Datensatz (K) verglichen werden.

Aufgabe 10 Uns stehen mittlerweile einige Möglichkeiten zur Verfügung. Die Nullhypothese soll sein, dass die Behandlung der Wolken keine Auswirkungen hat; sie also aus der selben Verteilung stammen.

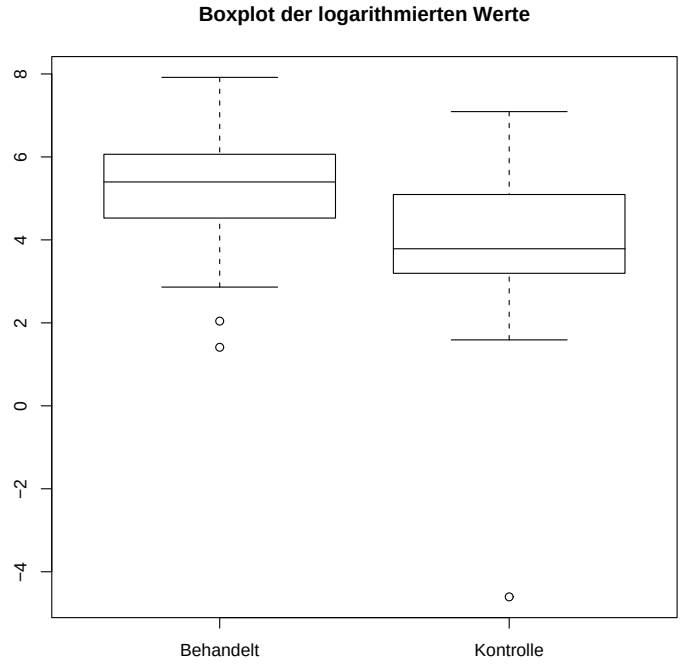
Sehen wir uns ersteinmal den QQPlot an



Man sieht, dass die behandelten Wolken schon zu etwas mehr Niederschlag neigen. Um die Ausreisser etwas runterzudrücken sehen wir uns dasselbe nochmal lagarithmiert an:



Die behandelten Wolken bleiben schön gleichmäßig und unterhalb der Diagonalen, was die Vermutung einer Verschiebung, hin zu mehr Niederschlag unterstützt.
Um ein noch besseres Bild zu bekommen, sehen wir uns den dazugehörigen geglätteten Boxplot an:



Hier ist die Verschiebung auch gut zu erkennen. Ein Blick auf die Werte, die *boxplot.stats* liefert zeigt, dass die lower whisker/hinge, mean und median um etwa 1.4

verschoben sind (upper whisker/hinge nur um 0.9), was immerhin einen Faktor von $\exp(1.4) \approx 4$ ausmacht (die 3 ausfallenden Werte betrachten wir als Ausnahmen). Zu guter letzt liefert uns noch der Wilcoxon-Test[2] einen p-Wert von 0.013. Da macht es auch nichts, dass uns Tukeys Test kein brauchbares Ergebnis liefert ($k_1 + k_2 = 4$).
Fazit: Behandelt fördert Regen in gewissem Maße.

Aufgabe 11 X ist ein Schätzer für einen Wert ϑ zum Konfidenzintervall $\frac{2}{3}$. Das bedeutet, das für ein Experiment $x_1, x_2, \ldots, x_n$ der Schätzer nach einer gewissen Regel f einen Wert liefert $f(x_1, \ldots, x_n) = x$, so dass $P(\{\vartheta \in [\alpha x, \beta x]\}) = \frac{2}{3}$. Da das kleine x ja für eine realisierte Zufallsvariable steht, ist jedes X_i eine Zufallsvariable bezüglich eines Experiments.
M_n ist nun der Stichprobenmedian der $X_1, \ldots, X_n$. Wann liegt nun ϑ nicht im Intervall $[\alpha M_n, \beta M_n]$? Genau dann, wenn mehr als die Hälfte aller αX_i rechts daneben respektive mehr als die Hälfte aller βX_i links neben ϑ liegen:

$$\vartheta \notin [\alpha M_n, \beta M_n] \Leftrightarrow \left(\#\{i : \alpha X_i > \vartheta\} > \frac{n}{2} \vee \#\{i : \beta X_i < \vartheta\} > \frac{n}{2}\right).$$

Nun sind alle X_i iid und

$$P(\alpha X \leq \vartheta \leq \beta X) \geq \frac{2}{3} \Rightarrow P(\vartheta < \alpha X) < \frac{1}{3} \text{ und } P(\vartheta > \beta X) < \frac{1}{3}.$$

Sehen wir uns die linke Seite des Konfidenzintervalls für $n \to \infty$ an: ϑ liegt genau dann links von αM_n, wenn von n Versuchen mehr als $\frac{n}{2}$ mit einer Wahrscheinlichkeit von $\frac{1}{3}$ erfolgreich waren. Genau hier greift nun aber das angegebene Resultat über die grossen Abweichungen beim Münzwurf:
Im Erwartungswert liegen $\frac{n}{3}$ der ϑ's links von αX. Um aber links von αM_n zu liegen, muss das für mehr als $\frac{n}{2}$ gelten. Die Wahrscheinlichkeit, dass dies eintritt konvergiert expoenentiell schnell gegen Null.
Dasselbe gilt für die rechte Seite von J_n und da die Summe von zwei exponentiell schnell abfallenden Folgen ebenfalls exponentiell schnell abfällt, ist alles gezeigt.

Etwas genauer formuliert:

$$\vartheta < \alpha M_n \Leftrightarrow S_n \geq \frac{n}{2} \quad Y_i = \begin{cases} 1 & \text{wenn } \vartheta < \alpha X, \\ 0 & \text{sonst.} \end{cases}, p_0 := P(Y = 1) \leq \frac{1}{3}$$

und somit

$$P(\vartheta < \alpha M_n) = P_{p_0}(\sum_{i=1}^{n} Y_i \geq \frac{n}{2}) \leq c^n \text{ für ein } c < 1.$$

Aufgabe 12
a)
Zum initialisieren der Daten:
`library(MASS) data(galaxies) galax<-galaxies/1000`

1. Der Median ist der Mittelwert M zwischen dem 41. und 42. Wert aus der sortierten Liste `sort(galax)` Wir wollen nun zwei Werte a, b wissen, so dass für eine 82-malige Ziehung mit zurücklegen $\tilde{X}$ aus der vorliegenden Stichprobe `galax` folgendes gilt:

$$P(\text{median}(\tilde{X}) < a) \leq 0.025 \text{ und } P(\text{median}(\tilde{X}) \leq b) \geq 0.975.$$

[2] ob nun mit den logarithmierten Werten oder nicht spielt keine Rolle, da der Rangsummentest gegenüber monotonen transformation invariant ist

Wir suchen also das 0.025-Quantil der Binomialverteilung bei 82 Zügen[3]: `qbinom`$(0.025, 82, 0.5) = 32$ und aus Symmetriegründen die obere Grenze `qbinom`$(0.975, 82, 0.5) = 50$.

Der 32. und der 51. Wert aus `sort(galax)` liefern das 95% Konfidenzintervall $2.0166, 2.196$

2. `bootinfo<-boot(galax,function(y,j) median(y[j]),R=1000)`

   `boot.ci(bootinfo,type=''basic'')`

   Liefert als ein 95% Konfidenzintervall: $1.973, 2.150$.

3. `bootinfo<-boot(galax,function(y,j) c(median(y[j]),sd(y[j])),R=1000)`

   `boot.ci(bootinfo,type=''stud'')`

   Liefert als ein 95% Konfidenzintervall: $1.952, 2.153$.

b)

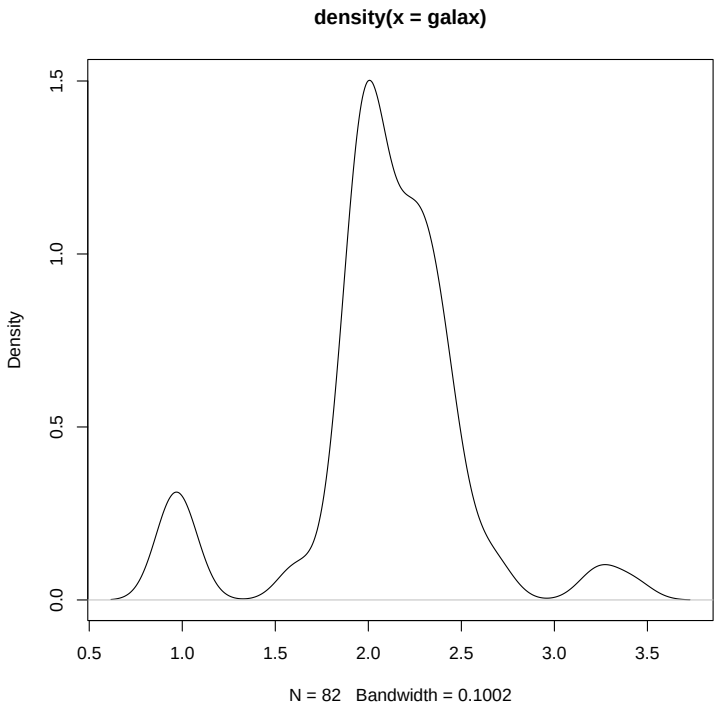
Ein Aufruf des Programms zur Aufgabe 12 bringt (nach einer ganzen Weile Rechnerei) folgende Resultate
Mit `source(''Aufgabe12.r'')` kommt heraus

exakt Beim Exakten Test lag er 917 mal richtig.

basic Beim basic-bootstrap lag er 808 mal richtig.

stud Beim studentisierten-bootstrap lag er 815 mal richtig.

Dies geht auch einigermaßen mit unseren Erwartungen konform: der exakte Wert ist der beste, mit Abstand gefolgt vom studentisierten- und dann der basic-Bootstrap. Um ein Bild von der Verteilungen zu bekommen, kann man sich das dazugehörige Bildchen anschauen:



[3]siehe auch die Vorlesung vom 1.11.01