

Prof. Dr. Antony Unwin, Alexander Pilhöfer
Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Institut für Mathematik
Universität Augsburg
<http://stats.math.uni-augsburg.de/>

Stochastik für Lehramt

Übungsblatt 9

Abgabe: Montag 28. Juni 2010, bis spätestens 10.00 Uhr; Briefkasten: Stochastik für Lehramt

Die Aufgaben können auch in Gruppen bearbeitet und abgegeben werden! Für jede Aufgabe gibt es 5 Punkte. Lösungen in R können in Form eines Skriptes (Textdatei) per email an die jeweiligen Übungsgruppenleiter geschickt werden!

Grundlegendes:

- (a) Was sind *Varianz*, *Bias* und *MSE*?
- (b) Wie hängen die drei o.g. Größen zusammen?
- (c) Was ist ein Punktschätzer? Was ist daran nachteilig?
- (d) Gegen welche Verteilung konvergiert $X \sim B(N, p)$ für $N \rightarrow \infty$?

1. (MONDRIAN)

Laden Sie den Datensatz `EcoTestAutos2009` in `Mondrian`.

- (a) Was beschreibt die Variable *DPF*?
- (b) Erstellen Sie Scatterplots von den stetigen Variablen. Gibt es Ausreißer, und wenn ja, welche? Verwenden Sie zur Identifizierung
 - i. Highlighting
 - ii. Abfragen: mit `STRG+SHIFT+Mouse-Over` werden die im Hauptfenster angewählten Variablen angezeigt.
- (c) Wie sind die Variablen *Punkte.CO2*, *Punkte.Schadstoffe*, *Gesamt* und *Sterne* zu interpretieren und wie hängen Sie zusammen?
- (d) In einem Scatterplot von *Verbrauch* und *Punkte.CO2* zeigt sich ein sehr auffälliges Muster mit mehreren Gruppen. Was sind diese Gruppen und wie kommt es vermutlich zu diesem Muster?

2. Simulieren Sie in R mit der Funktion `replicate()` $m = 1000$ Stichproben des Umfangs $n = 625$ aus einer Verteilung ihrer Wahl.

- (a) Berechnen Sie für jede Stichprobe den Mittelwert.
- (b) Berechnen Sie für jede Stichprobe ein Konfidenzintervall für den Erwartungswert.
- (c) Geben Sie die Anzahl der Stichproben an, für die der Mittelwert außerhalb des Konfidenzintervalls liegt.
- (d) Was ist ein Konfidenzintervall?
- (e) Wie ändert sich die Anzahl in Aufgabe b), wenn man m erhöht? Wie, wenn man n erhöht?

- (f) (Zusatz, schwierig) Sortieren Sie die Stichproben nach ihrem Mittelwert. Plotten Sie die Mittelwerte als Punkte. Fügen die Intervalle (z.B. mit Hilfe der Funktion `lines(x = c(i, i), y = c(mini, maxi))`) hinzu. Verwenden Sie eine andere Farbe, falls für diese Stichprobe der Erwartungswert außerhalb des CIs liegt. Zeichnen Sie den Erwartungswert als Linie ein.
3. Seien X_1, X_2, \dots, X_n unabhängige, identisch verteilte Zufallsvariablen, deren Verteilung den Erwartungswert μ und die Varianz σ^2 besitzt.

Bearbeiten Sie die nachfolgenden Teilaufgaben, indem Sie hierbei jeden Schritt angeben und begründen.

- (a) Zeigen Sie, dass das arithmetische Mittel $\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i$ ein erwartungstreuer Schätzer für μ ist. Berechnen Sie den mittleren quadratischen Fehler dieses Schätzers.
- (b) Zeigen Sie, dass die Stichprobenvarianz $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ ein erwartungstreuer Schätzer für σ^2 ist. Berechnen Sie den Bias des Schätzers $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ für σ^2 .
- (c) Ist μ bekannt, so ist $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ ein erwartungstreuer Schätzer für σ^2 .
- (d) (Zusatz) Ist $(\frac{1}{n-1} \sum_{i=1}^n X_i^2) - \frac{n}{n-1} \bar{X}^2$ ein erwartungstreuer und konsistenter Schätzer für σ^2 ?
4. Eine Online-Umfrage zum Thema Abtreibungspille 'Mifegyne' ergab, dass von 4,558 Befragten 1,102 die Einführung dieser Pille gut fanden, wohingegen 3,456 dies ablehnten (Quelle: <http://www.pro-leben.de/abtr/umfragen.php>).

- (a) Diskutieren Sie, ob obiger Sachverhalt wie folgt formal gefasst werden darf: Eine Zufallsstichprobe x_1, x_2, \dots, x_n —unabhängig, identisch verteilt—stammt aus einer Bernoulli-Verteilung ($x \in \{0, 1\}$):

$$f(x; p) = p^x q^{(1-x)} \quad p + q = 1, p, q > 0.$$

Spezifizieren Sie hierbei explizit die Interpretationen aller mathematischen Größen des obigen Modells im Anwendungskontext der Abtreibungsumfrage.

- (b) Welchen Schätzer \hat{p} würden Sie intuitiv für p vorschlagen?
- (c) Plotten Sie in \mathbb{R} die entsprechende Likelihood-Funktion und machen Sie sich ein Bild über den Maximum-Likelihood-Schätzer.

Bestimmen Sie nach den drei in der Vorlesung besprochenen Konfidenzschätzverfahren

- 'normal approximativ',
- 'normal plug-in',
- 'normal genau'

konkrete Schätzintervalle für den Proportionsparameter p der Binomialverteilung aus dem Umfragebeispiel zum Konfidenzniveau 0.99. Diskutieren Sie Ihre Ergebnisse.

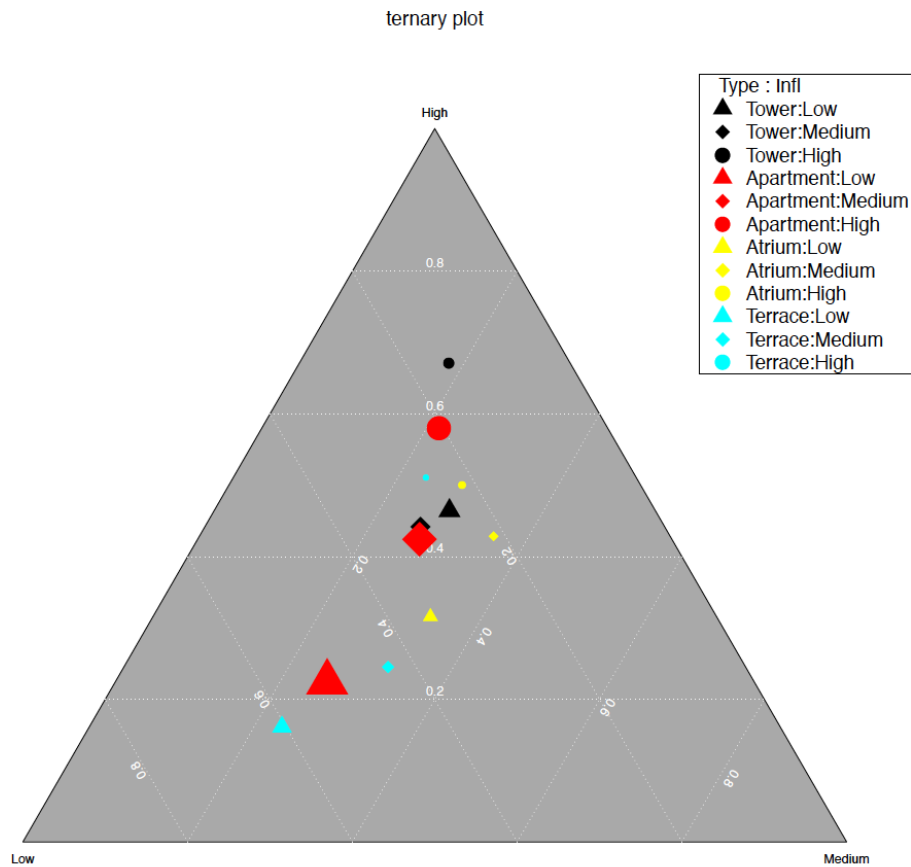


Abbildung 1: Ternaryplot des *Copenhagen* Datensatzes erstellt mit dem R-Paket *vcd*.

5. Untersuchen Sie die folgende Grafik zum Datensatz *Copenhagen*, der auch auf der Vorlesungswebseite verfügbar ist. Die Variablen des Datensatzes sind:

- Sat Gesamtzufriedenheit mit der Wohnung
- Infl Möglichkeit der Einflussnahme auf die Wohnbedingungen
- Type Art der Wohnung
- Cont Kontakt zu anderen Bewohnern (nicht in der Grafik)

- (a) Beurteilen Sie die Grafik in Bezug auf Suggestivität, Eindeutigkeit und Ästhetik.
- (b) Was wird dargestellt? Welche Variablen fließen in welcher Weise in die Grafik ein? Welche Größen werden explizit dargestellt?
- (c) Welche Hauptaussagen würden Sie basierend auf dieser Grafik treffen?
- (d) Wie hätten Sie selbst die Daten dargestellt?

-
- R
- (a) Machen Sie 100 Münzwürfe mit der Funktion `sample`.
 - (b) Machen Sie 100 Münzwürfe mit den Funktionen `runif` und `round`.
 - (c) Würfeln Sie 100 Mal (1-6) mit den Funktionen `runif` und `ceiling`.