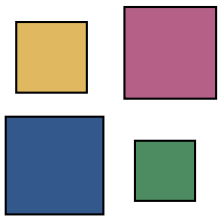

Prof. Antony Unwin, Alexander Pilhöfer
Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Institut für Mathematik
Universität Augsburg
http://stats.math.uni-augsburg.de/

Statistik I
Übungsblatt 5

Abgabe: *Dienstag 22. Mai 2012, bis spätestens 12.00 Uhr*; Briefkasten: Statistik I oder per email an die Übungsleiter

Die Aufgaben können auch in 2er-Gruppen bearbeitet und abgegeben werden!

1. (a) Welche Formel liegt dem Befehl `cor(x, y)` zugrunde? **(1P)**
   (b) Zeigen Sie, dass $MSE = Varianz + Bias^2$! **(1P)**
   (c) Worauf weist ein breites Konfidenzintervall hin? **(1P)**
   (d) Kann das Ergebnis eines Fußballmatches durch `rpois(1, lambda1) - rpois(1, lambda2)` sinnvoll simuliert werden? **(1P)**

2. **Chancentod (5P)**

   Auf der Internetseite `http://www.wahretabelle.de/wahretabelle/chancentod.php` finden Sie eine Tabelle zur Chancenverwertung der Bundesligateams der vergangen Saison. Untersuchen Sie in folgender Weise in `R`, ob sich die Trefferwahrscheinlichkeit für den 1. FC Kaiserslautern von den anderen Teams unterscheidet:

   (a) Schätzen Sie für die anderen Teams eine mittlere Trefferrate. Berechnen Sie den symmetrischen Annahmebereich für die Anzahl der Tore, wenn Sie von einem Signifikanzniveau von 0.05 bzw. 0.01 und einer Wahrscheinlichkeit von p gleich dieser mittleren Trefferrate ausgehen. Wäre die Trefferwahrscheinlichkeit des 1. FC Kaiserslautern jeweils signifikant verschieden?

   (b) Bestimmen Sie mittels des Befehls `binom.test()` den p-Wert. Lassen Sie sich hierzu einmal ein 95%-Konfidenzintervall und einmal ein 99%-Konfidenzintervall ausgeben.

   (c) Nehmen Sie an, die Trefferwahrscheinlichkeit p einer Mannschaft sei vor der Saison völlig unbekannt. Sie sollen diese Wahrscheinlichkeit nach einigen Spielen für die Mannschaft schätzen. Wieviele Chancen muss sich die Mannschaft herausspielen, damit ein 95%-Konfidenzintervall für alle möglichen wahren Werte p nicht länger als 2 Prozent ist?

   (d) Es sollen statistische Tests mit dem geschätzten Wert $\hat{p}$ gemacht werden. Die Nullhypothese hat die Form $H_0 : p = p_0$. Wann ist es sinnvoll, das Konfidenzintervall (bzw. den Annahmebereich unter H_0) besonders klein zu machen? Was bedeutet dies für die Stichprobengröße?

3. **KI (5P)**

   Simulieren Sie in `R` 100-mal $1,000$ Realisationen einer Normalverteilung mit den Parametern $\mu = 2$ und $\sigma = 5$. (der Befehl `replicate` könnte hilfreich sein) Bestimmen Sie in jedem dieser 100 Fälle das Konfidenzintervall für μ zum Konfidenzniveau 0.95 bei bekannt vorausgesetztem σ^2. Bestätigt sich die Aussage *'Wenn wir viele Stichproben ziehen und Konfidenzintervalle aus jeder berechnen, werden wir finden, dass ungefähr 95% der Intervalle den wahren Wert enthalten werden.'*? Stellen Sie das Ganze in geeigneter Weise graphisch dar.

4. **Flight Delay (5P)**

   Laden Sie den Datensatz *airATL2007* von der Webseite der Vorlesung herunter. Dieser enthält Daten zu sämtlichen Flügen, die im Jahr 2007 vom Flughafen *ATL* (Atlanta) aus zu Zielen innerhalb der USA gestartet sind. Eine Beschreibung der Variablen finden Sie unter:

   stat-computing.org/dataexpo/2009/the-data.html

   (a) Verschaffen Sie sich einen Überblick über den Datensatz. Geben Sie die wichtigsten Eckpunkte der Daten an.

   (b) War die durchschnittliche Verspätung der verspäteten Flüge höher während der Woche oder am Wochenende?

   (c) Betrachten Sie *ArrTime* gegen *DepTime*. Was fällt auf und woran liegt das? Kann man dieses Problem beheben?

   (d) Überprüfen Sie die Konsistenz der Daten im Hinblick auf die angegebenen Verspätungen: lassen sich die Verspätungen aus den angegebenen Abflug- und Ankunftszeiten rekonstruieren?

   (e) Welche fünf Airlines starten am häufigsten von diesem Flughafen? Mit welcher Airline sollte man fliegen, um möglichst wenig Verspätung zu haben?

5. **UK energy (5P)**

   Betrachten Sie die Grafik zum Thema "energy consumption in the UK' unter folgendem link:

   http://www.guardian.co.uk/news/datablog/2012/apr/27/
   uk-energy-sources-graphic#zoomed-picture

   (a) Was genau wird dargestellt? Welche Größen fließen ein und wie sind sie visualisiert?

   (b) Beurteilen Sie die Grafik im Hinblick auf Übersichtlichkeit, Suggestivität und Ästhetik.

   (c) Welche Aussagen können Ihrer Meinung nach auf Basis der Grafik getroffen werden?

   (d) Welche Aspekte der Daten sind interessant und sind diese optimal visualisiert?

   (e) Geben Sie mindestens eine alternative Visualisierungsform an, mit der Sie (mindestens) einen der Aspekte aus voriger Teilaufgabe besser datstellen können!