Übungen zur Wahrscheinlichkeitsrechnung und Statistik für Biologen

Blatt 4

---

1. Aufgabe Entgegen der landläufigen Meinung gibt es nicht nur grüne Marsmenschen, sondern auch rote und blaue. In der Tabelle `mars.txt` finden Sie die Größen (in cm) und Färbungen von allen 42 Marsbewohnern, die in den letzten 50 Jahren in verschiedenen havarierten Raumschiffen gefunden wurden. Wir gehen einmal davon aus, dass es sich dabei bezüglich der Größen um repräsentative Stichproben erwachsener Marsbewohner der verschiedenen Färbungen handelt.

1. Visualisieren Sie die Daten als Histogramm, als Boxplot und als Streudiagramm. Der R-Befehl für das Streudiagramm ist hier

```
> stripchart(size~color, method="jitter", pch=16,
  col=c("blue","green","red"), ylim=c(0.5,3.5))
```

   Dabei rüttelt „jitter“ die Punkte und `ylim=` setzt den Bereich der y-Achse. Vergleichen Sie die Vor- und Nachteile von Histogramm, Boxplot und Streudiagramm

2. Schätzen Sie aus den Daten die durchschnittlichen Größen roter, blauer und grüner Marsbewohner.

3. Wie genau sind diese Schätzungen? Berechnen Sie die Standardfehler!

4. Visualisieren Sie die Daten zusammen mit den Schätzungen und den Standardfehlern.

2. Aufgabe Ein leidenschaftlicher Risiko-Spieler Max hat die letzten Spiele verloren, da er mit seinem Lieblingswürfel zu kleine Werte gewürfelt hat. Nun zweifelt er an seinem Würfel und möchte testen, ob wirklich im Mittel jedes sechste Mal die 6 kommt. Dazu würfelt er 120 mal und notiert die jeweilige Augenzahl. Es kam 12 mal die 6 und sonst andere Zahlen. Da nur jedes zehnte Mal die 6 kam, fühlt sich Max bestätigt, möchte seine Vermutung nun aber auch statistisch belegen. Bitte helfen Sie ihm dabei. Hinweis: Max möchte zeigen, dass sein Würfel nicht fair ist. Ein Skeptiker würde entgegnen, dass die Beobachtung reiner Zufall ist. Wie groß ist unter dieser Nullhypothese die Wahrscheinlichkeit, eine so große oder noch größere Abweichung vom Erwartungswert zu sehen? Sie müssen sich zunächst die Verteilung unter der Nullhypothese überlegen.

3. Aufgabe Es mag überraschen, dass bei der Stichprobenvarianz durch die Stichprobenlänge minus eins geteilt wird. Wir wollen dies nun untersuchen.
In Aufgabe 3 des 3. Blattes wurde die Varianz des Würfels berechnet mit dem Ergebnis 35/12. Um Vertrauen zur Skalierung '1 / Stichprobenlänge minus eins' zu gewinnen, wollen wir diese Varianz aus Stichproben schätzen.

1. Generieren Sie eine Reihe von 20 Würfelwürfen. Verwenden Sie dazu den R-Befehl `sample()` mit der Option `replace=TRUE`. Schätzen Sie aus dieser Stichprobe die Varianz des Würfel durch die Stichprobenvarianz. Die Stichprobenvarianz der Stichprobe sei mit s_1 bezeichnet.

2. Wiederholen Sie diese Schätzung 1000 mal, d.h. samplen Sie 1000 mal eine Würfelreihe der Länge 20. Speicheren Sie die zugehörigen 1000 Stichprobenvarianzen $s_1, \ldots, s_{1000}$ in einen Vektor.

3. Berechnen Sie nun den Mittelwert dieser 1000 Stichprobenvarianzen. Vergleichen Sie das Resultat mit 35/12.

4. Wiederholen Sie obige Schritte mit der Skalierung '1 / Stichprobenlänge'. Vergleichen Sie das Resultat ebenfalls mit 35/12.

4. Aufgabe Um unser Gefühl für Korrelationen zu verbessern, wollen wir uns Datenwolken zu gegebenen Korrelationen überlegen. Dabei gibt es keine eindeutige Lösungen; verschieden Punktwolken passen zu den Korrelationen. Zeichnen Sie jeweils eine Punktwolke mit 8 Punkten.

1. Wie könnte eine Punktwolke mit Korrelation 0.9 aussehen?

2. Wie könnte eine Punktwolke mit Korrelation -0.4 aussehen?

3. Wie könnte eine Punktwolke mit Korrelation -1 aussehen?

4. Wie könnte eine Punktwolke mit Korrelation 0.01 aussehen?

5. Wie könnte eine Punktwolke mit Korrelation -0.85 aussehen?

Noch ein reales Datenbeispiel: Führen Sie den R-Befehl `> data(trees); attach(trees)` aus. Eine Beschreibung des Datensatzes erhalten Sie mit `> ?trees`. Berechnen Sie die Korrelation zwischen 'Volume' und 'Girth' mit dem R-Befehl `cor()`. Wie könnte die zugehörige Punktwolke aussehen? Bestätigen Sie Ihre Vermutung, indem Sie 'Volume' in Abhängigkeit von 'Girth' plotten.

5. Aufgabe Erzeugen Sie sich mit den folgenden Befehlen eine Gesamtpopulation V von reellen Werten.

```
N <- 100000
V <- rbeta(N,0.5,0.5)*100
```

1. Visualisieren Sie die Verteilung in angemessener Weise und zeichnen Sie auch Mittelwert und Standardabweichung ein.

2. Ziehen Sie eine Stichprobe vom Umfang $n = 20$ aus der Gesamtpopulation und berechnen Sie Mittelwert und Standardfehler.

3. Ziehen Sie nun aus derselben Population V 1000 unabhängige Stichproben vom jeweiligen Umfang $n = 20$ und untersuchen Sie wie häufig der Stichprobenmittelwert mehr als eine bzw. mehr als zwei Standardfehler vom tatsächlichen Mittelwert von V entfernt ist. Folgender R-Code kann Ihnen dabei nützlich sein:

```
M <- numeric(1000)      # Vektor fuer die 1000 Stichprobenmittelwerte
S <- numeric(1000)      # Vektor fuer die 1000 Standardfehler
n <- 20                 # Stichprobenlaenge
for(i in 1:1000) {
   s <- sample(???)     # Sample aus dem Vektor V 20 Werte
   M[i] <- ???          # Berechne den Mittelwert der Stichprobe
   S[i] <- ???          # Berechne den Standardfehler der Stichprobe
}
sum( abs(M-mean(V))>S )
```

Überlegen Sie selbst, durch was Sie die Fragezeichen ersetzen müssen und welche weitere Zeile Sie benötigen, um auch den Fall mit zwei Standardfehlern zu behandeln. Wieso passt das Ergebnis zur '2/3-Faustregel' aus der Vorlesung?

4. Führen sie den vorherigen Aufgabenteil auch mit den Stichprobengrößen 5 und 50 durch. Was fällt Ihnen auf?

6. Aufgabe Erzeugen Sie sich mit den folgenden Befehlen eine zufällige Verteilung V:

```
N <- 100000
V <- rbeta(N,0.5,0.5)*100+rgamma(N,20,0.5)
```

1. Visualisieren Sie die Verteilung in angemessener Weise und zeichnen Sie Mittelwert und Standardabweichung ein.

2. Ziehen Sie 1000 Stichproben, jeweils vom Umfang $n = 20$ aus V und berechnen Sie jeweils den Stichprobenmittelwert.

3. Erstellen Sie ein Dichte-Histogramm der 1000 Stichprobenmittelwerte.

4. Berechnen Sie für die 1000 Stichprobenmittelwerte den Mittelwert m Standardabweichung s und fügen Sie zum Dichtehistogramm die Glockenkurve der Normalverteilung mit Mittelwert m und Standardabweichung s hinzu. Tipp: versuchen Sie es mit Befehlen der Art:

   ```
   x <- seq(from=???,to=???,by=???)
   lines(x,dnorm(x,m,s))
   ```

   (Probieren Sie aus, durch welche Zahlenwerte Sie die Fragezeichen ersetzen müssen.)

5. Wiederholen Sie die vorherigen Aufgabenteil auch für die Stichprobengrößen 5 und 50. Was fällt Ihnen auf?