Übung 2

1. Auf einer WWF-Schiffsexpedition werden an 50 aufeinanderfolgenden Tagen die Anzahlen x_i, $i = 1, 2, \ldots, 50$ der von weitem beobachteten Albatrosse (grosse Hochseevögel) aufgezeichnet:

```
1 0 1 2 0 2 1 0 1 1 0 3 1 0 1 1 0 1 4 0 0 1 0 1 0
0 0 2 1 1 1 2 0 2 0 3 3 0 0 0 1 0 1 1 2 0 0 1 0 2
```

a) Zeichne ein Stabdiagramm.

b) Sei n_k die Anzahl der Tage, an denen k Albatrosse beobachtet werden; die Grösse n_k/n heisst *relative Häufigkeit* der Tage mit k beobachteten Albatrossen. Berechne den Mittelwert $\overline{x}$ des obigen Datensatzes.
(Überlege, dass der Mittelwert bei Daten mit einer Klasseneinteilung auch mit der Formel $\overline{x} = \sum_k k \cdot n_k/n$ berechnet werden kann. Diese Grösse lässt sich hier einfacher berechnen als die bekannte Formel $\overline{x} = 1/n \sum_{i=1}^{n} x_i$.).

c) Wir wollen an die Daten eine Poisson-Verteilung anpassen, wobei der unbekannte Parameter λ durch $\overline{x}$ geschätzt wird.
Ergänze dazu die folgende Tabelle:

Anzahl der Albatrosse, k	beobachtete Anzahl der Tage, n_k	erwartete Anzahl der Tage, $n\widehat{p}_k$	$(n_k - n\widehat{p}_k)^2/n\widehat{p}_k$
0	21	20.33	0.022
$\vdots$	$\vdots$	$\vdots$	$\vdots$
≥ 5	0	$\ldots$	$\ldots$

d) Passt das Modell gut zu den Daten ? (Überschlag, mit der Faustregel)

2. Im Rahmen eines Experimentes zur Bestimmung der Abbaufähigkeit eines Bodens wurde die maximale Dehnung von Kunststoffstreifen gemessen. Die Messung ergab folgende Resultate :

2.1 4.4 2.0 2.9 2.9 3.2 2.9 3.1 3.5 2.4 3.2

a) Zeichne eine Stamm-und-Blatt-Darstellung (stem and leaf display) der Daten.

b) Zeichne die empirische kumulative Verteilungsfunktion der Daten.
Wo übersteigt die Verteilungsfunktion 25%? Wo 90%?

c) Bestimme aus a) oder b) den Median, das obere und das untere Quartil und (freiwillig) zeichne einen Boxplot der Daten (vgl. Buch S. 26d,e)!

3. Eine Suspension (Lösung) von Bakterien (Keimen) wird auf 100 Petri-Schalen abgesetzt, pro Petri-Schale 0.01 ml. Die Keimzahl in der Suspension sei 1000/ml. Nach einer gewissen Zeit werden die Bakterienkolonien, von denen jede aus einem Keim entsteht, unter dem Mikroskop ausgezählt.

a) Wie gross ist λ_0, die *erwartete* Anzahl Bakterienkolonien pro Petri-Schale?

b) Welcher Verteilung sollte die Anzahl Bakterienkolonien pro Petri-Schale folgen? Begründe kurz!

c) Für einen Versuch braucht man Petri-Schalen, die nur genau eine Bakterienkolonie enthalten. Wie stark muss die Suspension verdünnt werden, damit die erwartete Anzahl solcher Schalen maximal ist? (Für welches λ ist die Wahrscheinlichkeit maximal, dass eine Schale *genau eine* Bakterienkolonie enthält?)

d) In einem anderen Fall möchte man die Suspension möglichst stark verdünnen, aber so, dass man erwarten kann, dass 95 der 100 Petri-Schalen mindestens eine Bakterienkolonie enthalten. Wie stark darf man maximal verdünnen? (Ab welchem λ ist die Wahrscheinlichkeit für eine einzelne Petri-Schale, *keine* Bakterienkolonien zu enthalten, $\leq 5\%$?)

e) Wie gross ist für $\lambda = 0.5$ die Wahrscheinlichkeit, dass eine nicht-leere Petri-Schale genau eine Bakterienkolonie enthält? (Salopp gefragt: Wieviele Prozent der Petri-Schalen, die *mindestens* eine Bakterienkolonie enthalten, werden *genau* eine enthalten?)

f) Wie klein muss λ sein, damit die Wahrscheinlichkeit $\frac{3}{4}$ ist, dass eine nicht-leere Petri-Schale genau eine Bakterienkolonie enthält?
Hinweis: Da die resultierende Gleichung für λ nicht geschlossen lösbar ist, reicht es, die Gleichung anzugeben. Wer Lust hat, kann noch die numerische Lösung systematisch oder durch "Pröbeln" bestimmen.

4. (**Anspruchsvolle** *R*-Aufgabe) Rutherford und Geiger haben die Emission der α-Teilchen einer radioaktiven Substanz gemessen. In der Tabelle bezeichnet n die Anzahl der α-Teilchen, die in einer Zeiteinheit ($= 7.5s$) beobachtet wurden; *frequency* gibt an, wie oft $n = 0, 1, \ldots, 14$ α-Teilchen beobachtet wurden (total 2612 Zeiteinheiten). (Die Daten befinden sich im Data Frame `alpha`.)

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
frequency	57	203	383	525	532	408	273	139	49	27	10	4	0	1	1	0

a) Warum eignet sich die Poissonverteilung zur Beschreibung dieser Daten?

b) Bestimme den Mittelwert der Daten. R-Anleitung: Das Data Frame `alpha` enthält die Variablen "n" und "frequency".

```
> alpha <- read.table.url("http://stat.ethz.ch/Teaching/Datasets/alpha.dat", header = T)  # Datensatz via Internet einlesen
> alpha                                                   # Daten anschauen
> nZeiteinheiten <- sum(alpha[,"frequency"])   # Anzahl Zeiteinheiten in denen gemessen wurde
> nAlpha <- ? (alpha[,"n"]*alpha[,"frequency"])          # gesamte Anzahl gemessener Alphateile
> mittlAlpha <- ? / ?                          # mittlere Anzahl Alphateile pro Zeiteinheit
```

c) Zeichne ein Stabdiagramm. Da Zeiteinheiten mit mehr als 11 Teilchen sehr selten auftreten, fassen wir die Zeiteinheiten mit 11 oder mehr Teilchen in eine Klasse 11+ zusammen. Dies hat sowohl grafische Gründe (Sichtbarkeit) als auch theoretische für den Anpassungstest (wird in der Vorlesung noch behandelt).

```
> motif()                                                 # Grafikfenster aufstarten
> alpha11 <- alpha[1:12,]                # neues Dataframe: für k=0,...,10, gleich wie alpha ...
> alpha11[12,"frequency"] <- sum(alpha[12:15,"frequency"])         # ... und letzte Zeile ≙ Klasse 11+
> plot(alpha11[,"n"], alpha11[,"frequency"], type = "h")           # Stabdiagramm zeichnen
```

d) Zeichne die Werte der Poissonverteilung mit Parameter λ in das Stabdiagramm ein, indem Du als Schätzer für den unbekannten Parameter λ den Mittelwert aus Unteraufgabe (b) verwendest.

```
> pk <- dpois(alpha11[,"n"], lambda=?)             # Vektor der W'keiten für k Alphateilchen
> pk[12] <- 1-ppois(10, lambda=?)                              # W'keit für Klasse 11+
> npkErw <- pk * ?                      # Erwartete Anzahl Zeiteinheiten mit k Alphateilchen
> points(alpha11[,"n"]+0.2, ?, pch="*")   # Punkte 0.2 Einheiten neben zugehörige Stäbe zeichnen
```

e) Ist die Poissonverteilung als Modell für diese Daten angebracht? Beurteile grafisch die Übereinstimmung der Daten mit der Poissonverteilung mit Hilfe der Faustregel für die einzelnen Klassen. Wir wollen zu diesem Zweck in unser bereits erstelltes Diagramm noch Linien zeichnen, die dem Intervall $n\widehat{p_k} \pm 2 * \sqrt{n\widehat{p_k}}$ entsprechen:

```
> for (i in 1:12)                       # führt nächsten Befehl für alle Zahlen von 1 bis 12 aus
> lines(rep(alpha11[i, "n"]+0.2, 2), npkErw[i] + 2 * sqrt(npkErw[i]) * c(-1,1))
```

\# *Zeichnet beim i-ten Stab eine vertikale Linie von* $n\widehat{p_k} - 2 * \sqrt{n\widehat{p_k}}$ *bis* $n\widehat{p_k} + 2 * \sqrt{n\widehat{p_k}}$

Abgabe: Montag, 10. April, in der Pause der Vorlesungsstunde