Musterlösung der Klausur vom 29. Juli 2003

Aufgabe 1.

|10

Definieren Sie die folgenden statistischen Begriffe in einem Satz oder in einer Formel:

1. Histogramm |1
   In einem Histogramm werden die Häufigkeiten unterschiedlicher Beobachtungen (bei diskreten Daten), bzw. Häufigkeiten von Beobachtungen in einem Intervall (bei kontinuierlichen Daten) nebeneinander als Rechtecke dargestellt, deren Flächen proportional zu den Häufigkeiten sind.

2. Empirische Varianz |1
   Die emp. Varianz ist die mittlere quadratische Abweichung einzelner Datenpunkte zum Mittelwert des ganzen Datensatzes. Oder als Formel:

$$Var_{\text{emp}}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 .$$

3. Varianz einer Zufallsvariablen |1
   Die (theoretische) Varianz einer Zufallsvariablen ist die erwartete quadratische Abweichung der Zufallsvariablen von ihrem Erwartungswert. Oder als Formel:

$$Var_{\text{theo}}(X) = E(X - EX)^2 .$$

4. Zentraler Grenzwertsatz |2
   Summen (Überlagerungen) vieler unabhängiger und identisch verteilter Zufallsvariablen sind normalverteilt. Satz von Moivre-Laplace: Seien $X_1, \ldots, X_n$ iid mit $EX_1 = \mu$ und $VarX_1 = \sigma^2$, und sei $S = \sum_{i=1}^{n} X_i$, dann gilt für $n \to \infty$:

$$P\left(\frac{S - n\mu}{\sqrt{n}\sigma}\right) \longrightarrow \int_{-\infty}^{t} \frac{1}{2\pi} e^{-\frac{1}{2}x^2}\, dx .$$

5. Binomial-Verteilung |1
   Die Binomial-Verteilung gibt die Wahrscheinlichkeit für k Erfolge bei n Versuchen an. Oder: Die Binomial-Verteilung ist die Verteilung einer Summe von n bernoulliverteilten Zufallsvariablen. Oder als Formel:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} .$$

6. Maximum-Likelihood-Methode |1
   In der Maximum-Likelihood-Methode werden die Parameter θ eines Modells M_θ so geschätzt, dass sie die Wahrscheinlichkeitsdichte der Daten gegeben das Modell (die Likelihood) maximieren. Als Formel bei i.i.d. verteilten Beobachtungen $x_1, \ldots, x_n$ mit Dichte $f_\theta(x)$:

$$\hat{\theta} \;=\; \operatorname*{argmax}_\theta Lik(D \mid M_\theta) \;=\; \operatorname*{argmax}_\theta \prod_{i=1}^{n} f_\theta(x_i) .$$

7. Overfitting | 3

Overfitting bedeutet, dass ein Modell (z.B. ein Klassifikator) nur gut auf die Daten passt, auf denen es trainiert worden ist, aber nicht auf unabhängige Testdaten. Allgemein: Das Modell lernt Stichproben-Eigenschaften und keine Populations-Eigenschaften.

Aufgabe 2. | 10

1. Die Zufallsvariable X nimmt die Zustände -2, -1, 0, 1, und 2 mit den Wahrscheinlichkeiten 0.3, 0.1, 0.2, 0.1 und 0.3 an.
   Berechnen Sie Erwartungswert und Varianz von X. | 1

   X ist symmetrisch um Null $\Rightarrow EX = 0$.

$$VarX = EX^2 - (EX)^2 = EX^2 = \sum_{i=-2}^{2} i^2 P(X=i) = 2\sum_{i=1}^{2} i^2 P(X=i)$$
$$= 2 \cdot 4 \cdot 0.3 + 2 \cdot 1 \cdot 0.1 = 2.4 + 0.2 = 2.6.$$

2. $Y \sim N(0,1)$ und $Z = 2 \cdot Y - 4$. Wie groß sind Erwartungswert, Median, Modalwert und Varianz von Z? | 1

   $EZ = E(2Y-4) = 2EY - 4 = 2 \cdot 0 - 4 = -4$.

   Y normalverteilt $\Rightarrow$ Z normalverteilt $\Rightarrow$ EZ = Median = Modalwert.

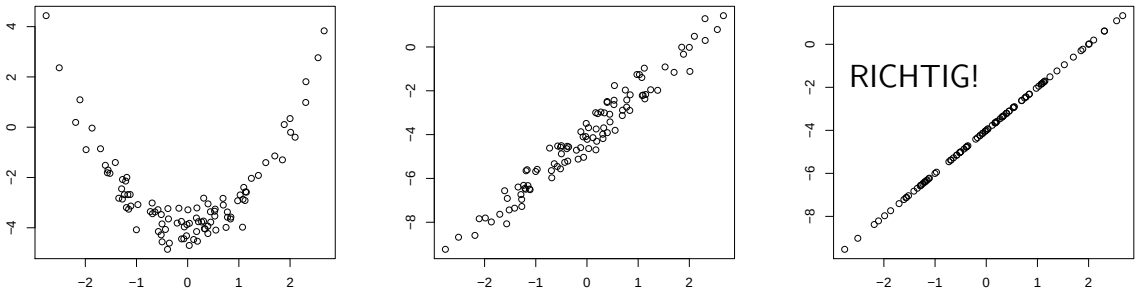
   $VarZ = Var(2Y-4) = Var(2Y) = 4VarY = 4$.

3. Sind Y und Z unabhängige Zufallsvariablen? Sind Y und Z unkorrelliert? | 1

   Z ist eine Lineartransformation von Y. Damit sind Y und Z deterministisch abhängig (und damit auch korrelliert).

4. Welcher der folgenden Scatterplots stellt mögliche Realisierungen der Zufallsvariablen Y und Z dar? Kreuzen Sie die richtige Graphik an. | 1

   Wenn Z eine Lineartransformation von Y ist, dann erhalten wir natürlich eine Gerade als Scatterplot.



5. Die Zufallsvariablen S und T seien unabhängig und normalverteilt ($N(\mu, \sigma^2)$) mit $S \sim N(0,4)$ und $T \sim N(0,1)$. Wie groß ist dann der Loglikelihood-Ratio von S verglichen mit T für den Wert 4? | 3

   Dichte der Normalverteilung $N(\mu, \sigma^2)$: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$.

$$LLR(x) = \log\left(\frac{f_S(x)}{f_T(x)}\right) = \log\left(\frac{\frac{1}{2}\exp\left\{-\frac{1}{2}\frac{x^2}{4}\right\}}{\exp\left\{-\frac{1}{2}x^2\right\}}\right) = \log\left(\frac{1}{2}\exp\left\{-\frac{1}{2}\frac{x^2}{4} + \frac{1}{2}x^2\right\}\right)$$
$$= \log\frac{1}{2} - \frac{1}{2}\frac{x^2}{4} + \frac{1}{2}x^2 \; = \frac{3}{8}x^2 - \log 2.$$
$$LLR(4) = \frac{3}{8}16 - \log 2 = 6 - \log 2 \approx 5.31.$$

6. Die Zufallsvariablen V und W nehmen gleichverteilt Werte in {A, C, G, T} an. Geben Sie **2 unterschiedliche Lösungen** für gemeinsame Verteilungen an, für die $P(V = W) = 0.5$ gilt. |3

   Schreibt man die gemeinsame Verteilung von V und W als (4×4)-Matrix, so muss diese Matrix folgende Bedingungen erfüllen:

   (a) V und W sind gleichverteilt. Also müssen sich die Zeilen und die Spalten jeweils zu 0.25 addieren, da die Zeilen- bzw Spaltensumme die Verteilungen von V und W darstellen.

   (b) Die Diagonalelemente müssen sich zu 0.5 addieren, damit gilt:

$$P(V = W) \; = \sum_{i\in\{\texttt{A,C,G,T}\}} P(V = i, W = i) \; = \; 0.5\;.$$

   Zwei simple Matrizen, die obige Bedingungen erfüllen, sind:

$$P(V,W) = \begin{pmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0.25 \\ 0 & 0 & 0.25 & 0 \end{pmatrix} \text{ und } \tilde{P}(V,W) = \begin{pmatrix} 0.1 & 0.15 & 0 & 0 \\ 0.15 & 0.1 & 0 & 0 \\ 0 & 0 & 0.15 & 0.1 \\ 0 & 0 & 0.1 & 0.15 \end{pmatrix}.$$

Aufgabe 3. |10

1. Kommentieren Sie die folgenden Zeilen R-Code. |3

```
M <- matrix(runif(100*1000,min=0,max=1),100,1000)
```

   - `runif(100*1000,min=0,max=1)` erzeugt einen Vektor der Länge 100 000 dessen Einträge unabhängige Realisierungen einer Zufallsvariablen $X \sim \mathsf{U}(0,1)$ (d.h. X ist gleichverteilt mit Minimum 0 und Maximum 1) darstellen.
   - `matrix( · , 100, 1000)` erzeugt eine Matrix mit 100 Zeilen und 1000 Spalten. Im obigen Fall wird sie mit den von `runif` erzeugten unabhängigen Realisierungen von X aufgefüllt.

   `M` ist also eine Matrix mit 100 Zeilen und 1000 Spalten. Die einzelnen Elemente von `M` sind unabhängige Realisierungen einer ZV $X \sim \mathsf{U}(0,1)$.

```
a <- apply(M,1,function(x){sum(x>0.99)})
```

   Dieser Befehl wendet auf die Matrix `M` zeilenweise die Funktion `f(x)` an, die auch gleich definiert wird. Als Resultat bekommt man einen Vektor `a` der Länge 100, in welchem für jede Zeile von `M` die Anzahl von Elementen gespeichert ist, die größer als 0.99 sind.

`b <- apply(M,1,mean)`
Der Vektor `b` hat die Länge 100. In ihm stehen die jeweiligen Zeilenmittel von `M`. Diese Zeile ist äquivalent zu `b <- rowMeans(M)`.

`c <- apply(M,2,max)`
In dieser Zeile wird die Funktion `max` spaltenweise auf `M` angewandt. `c` ist also ein Vektor der Länge 1000, in dem die Maxima der jeweiligen Spalten gespeichert sind.

2. Welche theoretische Verteilung hat `a`? Wie würden Sie den/die Parameter wählen? |3
Hier waren zwei Antworten möglich:

   - Jedes Element m_{ij} von `M` ist eine unabhängige Realisierung einer ZV $X \sim \mathrm{U}(0,1)$. Die Wahrscheinlichkeit, dass diese Realisierung größer als 0.99 ist, beträgt 1%. Jede Zeile ist also die Summe bernoulliverteilter Zufallsvariablen mit p =1%. Da es 1000 Summanden gibt, ist diese Summe binomialverteilt mit $p = 0.01$ und $n = 1000$.

     Da nun 1% eine nicht sonderlich große Erfolgswahrscheinlichkeit darstellt und n mit 1000 auch relativ groß ist, könnte man auch annehmen, dass man sich "auf dem Weg" zu einer Poissonverteilung mit $\lambda = n \cdot p = 10$ befindet.
   - Man kann vergessen, dass man die Verteilungsfunktion der m_{ij} kennt und den Parameter p der Binomialverteilung aus der Matrix `M` schätzen. Dafür würde man die relative Häufigkeit der Erfolge der unterliegenden Bernoulliexperimente verwenden, d.h.

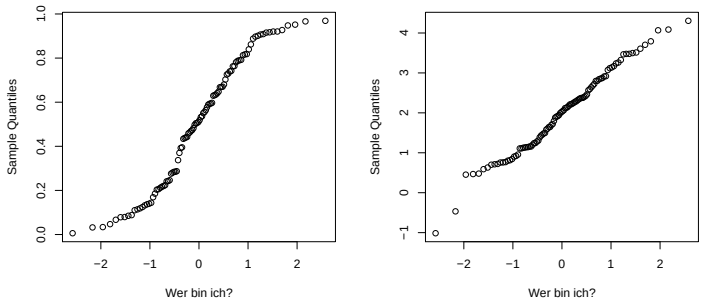
     ```
     p <- sum(M>0.99)/(dim(M)[1]*dim(M)[2])
     ```

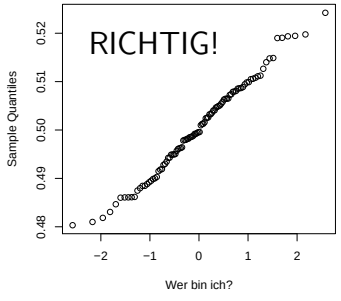
     Alternativ kann man wegen der seltenen Erfolge und der vielen Versuche auch auf eine Poissonverteilung schließen und für λ den Mittelwert von `a` als Maximum-Likelihood-Schätzer verwenden. Das liefert:

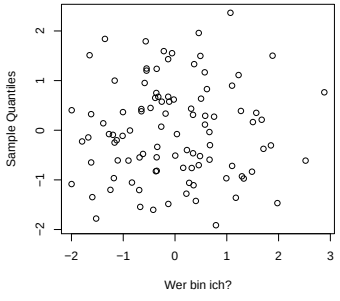
     ```
     l <- sum(a)/dim(M)[1]
     ```

     Bemerkung: Das obige `l` ergibt sich auch als n`p`, denn n=`dim(M)[2]` und `sum(M>0.99)` = `sum(a)`. Also gilt $n \cdot$ `p` = `l`.

3. Welcher der folgenden Plots ist das Ergebnis von `qqnorm(b)`? Kreuzen Sie die richtige Graphik an. |2



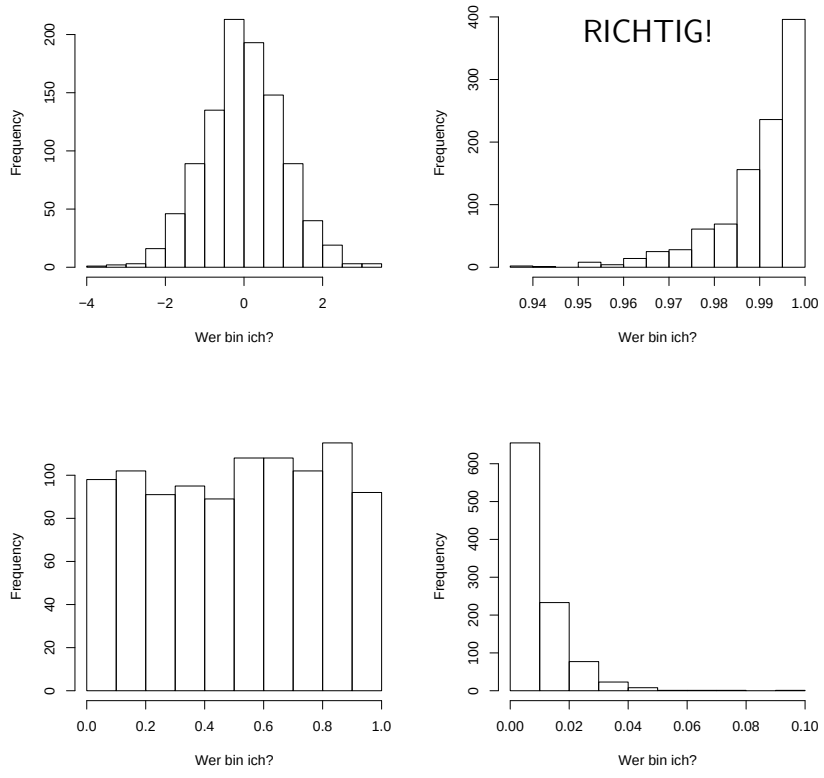




Gehen wir nach dem Ausschlussprinzip vor:

- Die in `b` gespeicherten Zeilenmittel liegen sicher im Intervall $[0, 1]$, worin dann auch die im QQ-Plot dargestellten Stichprobenquantile enthalten sind. Das schließt die rechten beiden Bilder aus.
- Die einzelnen Elemente in jeder Zeile von `M` stellen Realisierungen einer ZV $X \sim \mathrm{U}(0, 1)$ dar. Der Erwartungswert von X ist somit $1/2$. Die Zeilenmittel sind konsistente Schätzer dieses Erwartungswertes und sollten sich deshalb "in der Nähe" von $1/2$ aufhalten. Alle Stichprobenquantile sollten also "ca. $1/2$" sein. Somit ist der linke untere QQ-Plot der richtige.

4. Welcher der folgenden Plots ist das Ergebnis von `hist(c)`? Kreuzen Sie die richtige Graphik an. |2



Jeder Eintrag in `c` ist ein Spaltenmaximum von `M`. Er sollte also "in der Nähe" von 1 liegen. Dies ist aber nur für das rechte obere Histogramm der Fall. Deshalb ist dieses das richtige.

Aufgabe 4. |10

In einer Studie wird die Expression eines bestimmten Genes bei kranken und gesunden Patienten untersucht. Von jedem Patient liegen sowohl der Genexpressionswert als auch die Konzentration des resultierenden Proteins vor. Wie stellen Sie fest, welches Verfahren sich besser zur Diagnose eignet?

1. Diagnostische Marker. Wie lassen sich die Expressions- und Proteindaten eines Genes zur Diagnose nutzen? Erstmal für die Expressionsdaten, mit den Daten über Proteinkonzentration verfährt man analog. Wir schauen uns die Verteilung der Expressionswerte bei den Gesunden und den Kranken an. Danach wählen wir einen kritischen Wert c, mit dem wir die Daten eines neuen Patienten vergleichen. Nehmen wir an, unser Gen wird bei kranken Menschen überexprimiert, dann diagnostizieren wir alle zukünftigen Patienten als krank, bei denen der Expressionswert größer als c ist.

Wie finden wir einen kritischen Wert c? Zum Beispiel durch einen Vergleich der Likelihoods: Wir unterstellen bei Gesunden und Kranken eine Normalverteilung, schätzen Mittelwert und Varianz, und wählen c dort, wo die beiden Likelihoods gleich sind. Damit erhalten wir eine Entscheidungsregel D_{gen}, die den Expressionswert eines neuen Patienten mit diesem Wert c vergleicht. Analog erhält man für die Proteindaten eine Entscheidungsregel D_{prot}.

2. Evaluation. Wissen wir damit schon, welche der beiden Entscheidungsregeln besser ist als die andere? Nein! Wir kennen bis jetzt nur das Verhalten auf Trainingsdaten. Für die Diagnose ist aber die Generalisierungsfähigkeit auf zukünftige Daten wichtig. Diese schätzt man auf Testdaten, die von den Trainingsdaten unabhängig sind, oder durch Kreuzvalidierung, in der iterativ die Daten in Trainings- und Testdaten aufgespaltet werden.

3. Signifikanz. Im Schritt 2 werden wir feststellen, dass sich die Fehler für die beiden Entscheidungsregeln unterscheiden. Aber: sind diese Unterschiede zufälliges Rauschen oder ist die eine Entscheidungsregel systematisch besser als die andere? Das können wir so herausbekommen: Wir evaluieren unsere Entscheidungsfunktionen für mehrere Datensätze und testen danach mit dem t-Test, ob sich die Mittelwerte der Fehler signifikant voneinander unterscheiden.