Aufgabenblatt aus
Mathematischer Statistik

Bemerkung:
Die Aufgaben sind in 5 Bl"ocke unterteilt. Jeweils drei aufeinanderfolgende Beispiele bilden einen Block, also: (1,2,3), (4,5,6), usw. Jeder mu"s 5 Aufgaben ausarbeiten, und dabei jeweils eine aus jedem Block ausw"ahlen.

1. (Monte Carlo Experiment) Berechnen Sie die Kenngrößen $k_1, \ldots, k_{11}$ ($\bar{x}, \tilde{x}$ etc.) aus Tabelle 1.2 des Vorlesungsskriptums für die Variablen *fvc* und *alter* des Aimu Datensatzes.

   (a) Generieren Sie eine genügend große Anzahl von R Stichproben (z.B. $R = 2000$) aus einer Normalverteilung mit Erwartungswert $\bar{x}(fvc)$ und Varianz $s^2(fvc)$. Berechnen Sie für jede Stichprobe die Kennzahlen $k_1(r), \ldots, k_{11}(r)$, $r = 1, \ldots, R$.

   (b) Zeichnen Sie nun ein Histogramm für $k_l(r)$, $r = 1, \ldots, R$, für alle Kennzahlen k_l, $l = 1, \ldots, 11$. (Auf diese Weise erhalten sie eine Monte Carlo Approximation der Dichte von k_l unter Annahme von Normalverteilung mit Parameter $\bar{x}(fvc)$ und $s^2(fvc)$. Zeichen sie die Kenngrößen der Originalstichprobe (Aimu) in die Histogramme ein.

   (c) Wiederholen Sie dies für die Variable *Alter*.

   Geben Sie eine Interpretation der Ergebnisse.
   (*ftp://statistik.tu-graz.ac.at/angstat/data/aimu.****, $*** = txt$, für ein Ascii File, $*** = sav$ für ein SPSS File, $*** = sdd$ für ein SPLUS File; Login als anonymous.)

2. In der folgenden Tabelle sind die CPU-Zeiten (in Sekunden) von 25 Jobs, die an einem Großrechner ausgeführt wurden, aufgelistet.

1.17	1.61	1.16	1.38	3.53	1.23	3.76	1.94	0.96	4.75
0.15	2.41	0.71	0.02	1.59	0.19	0.82	0.47	2.16	2.01
0.92	0.75	2.59	3.07	1.40					

   (*ftp://statistik.tu-graz.ac.at/angstat/data/cpu.****,*txt*, *sav* und *sdd*.)

   - Berechnen Sie die Kenngrößen aus Tabelle 1.2 des Vorlesungsskriptums.
   - Erstellen Sie ein Histogramm mit geeigneter Klassenbreite, den Boxplot und den Symmetrieplot.
   - Fertigen Sie den QQ-Plot der oben angegebenen Beobachtungen versus die $N(0,1)$-Verteilung an. Was schließen Sie daraus?

3. Erzeugen Sie 10 Samples S_j, $j = 1, \ldots, 10$, der Größe $n = 20$ aus der Normalverteilung mit Parameter $\mu = 0$ und $\sigma^2 = 1$.

   (a) Berechnen Sie für jedes Sample die empirische Verteilungsfunktion und vergleichen Sie sie mit der $N(0,1)$-Verteilung (graphisch).

   (b) Variieren Sie die Parameter n, μ und σ^2.

4. Bei den Sonnenfleckendaten wurde über den Zeitraum von 1749-1977 die monatlichen Mittelwerte der Anzahl der Sonneflecken beobachtet. (1977 nur für die Monate Jänner bis März.) Die erste Spalte gibt die Mittlere Anzahl, die zweite den Code für das Beobachtungsjahr an (1=1749).
   - Definieren Sie eine neue Variable Jahr, welche das Bebachtungsjahr in der üblichen Zeitrechnung enthält.
   - Stellen Sie diese *Zeitreihe* mittels einer Serie von Boxplots dar (für jedes Jahr ein Boxplot, das geordnet nach Jahren und alles in einem Graphen).
   - Wiederholen Sie dies für den Zeitraum 1900-1977.

   (*ftp://statistik.tu-graz.ac.at/angstat/data/sunspots.txt*)

5. Generieren Sie 10 Samples S_j der Größe $n = 20, 100$ aus einer $N(0,1)$-Verteilung. Erstellen Sie den QQ-Plot für diese Samples, wobei Sie die $N(0,1)$ Verteilung als Referenzverteilung verwenden. Die auf diese Art und Weise erstellten QQ-Plots sollten Ihnen einen Eindruck vermitteln, wie ein QQ-Plot bei Zutreffen der Verteilungsannahme (nämlich: die Daten stammen aus einer $N(0,1)$-Verteilung!) aussehen sollte.

6. Generieren Sie 200 Samples S_j der Größe $n = 40$ aus einer $N(0,1)$-Verteilung. Berechnen Sie für jedes Sample S_j, $j = 1, \ldots, n$, den Median $\tilde{x}(j)$, dh. das 0.5-te Quantil (vgl. Vorlesungsskriptum) und das arithmetische Mittel $\bar{x}(j)$. Sie erhalten auf diese Weise zwei Vektoren $(\tilde{x}(1), \ldots, \tilde{x}(200))$, resp. $(\bar{x}(1), \ldots, \bar{x}(200))$, die nun mit den Methoden der beschreibenden Statistik untersucht werden sollen. Berechnen Sie Kenngrößen für Lage und Streuung. Erstellen Sie Boxplots und Histogramme für die beiden Vektoren. Kommentieren Sie Ihre Ergebnisse.

   Führen Sie diesselben Berechnungen mit der $U(-\sqrt{3}, \sqrt{3})$-Verteilung durch. Kommentieren Sie die Resultate!

7. Sei $X_1, \ldots, X_n \overset{iid}{\sim} Gamma(\alpha, \lambda)$, d.h. eine Stichprobe aus der Gammaverteilung mit Parameter α und λ mit Dichte

   $$f(x; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} I_{0,\infty}(x).$$

   (a) Leiten Sie die Momentenschätzer für α und λ her.

   (b) Sei $\mu = E(X_1)$. Reparametrisieren Sie die Dichtefunktion durch $\lambda = \frac{\alpha}{\mu}$, $\alpha = \alpha$. Wie lauten die Maximum Likelihood Gleichungen für (α, μ)?

   (c) Berechnen Sie die Fisher Informationsmatrix.

   (d) Berechnen Sie mittels der Newton Raphson Methode die Maximum Likelihood Schätzer für (α, μ) bei den CPU Daten. Wie lauten die Maximum Likelihood Schätzer für (α, λ)?

   (e) Werten Sie die Inverse der Fisher Informationsmatrix im Maximum Likelihood Schätzer aus. (Die Wurzeln aus den Diagonalelementen dieser Matrix liefern Schätzungen für die Standardabweichung der Maximum Likelihood Schätzungen – die sog. Standardfehler.)

8. Im folgenden wollen wir die Verteilung eines Momentenschätzers mit der Verteilung eines ML-Schätzers anhand eines Beispiels mittels Simulation vergleichen.

   Sei $X \sim U(0,\theta)$, dh. auf $[0,\theta]$ stetig gleichverteilt. Berechnen Sie zunächst den Momentenschätzer und den ML-Schätzer für θ.

   Nun generieren Sie m Stichproben der Größe n aus $U(0,\theta)$, $\theta = 1$, und berechnen für jedes die Momentenschätzung und die ML-Schätzung für θ. Die zwei resultierenden m dimensionalen Vektoren sind statististisch zu analysieren. Zeichnen Sie ein Histogramm, schätzen Sie Mittelwert und Bias, den Mittleren quadratischen Fehler und den Standardfehler der Schätzer. Führen Sie die Berechnungen für $m = 100$ resp. 500 und $n = 10$ resp. 100 durch.

   Fassen Sie die Ergebnisse Ihrer Analyse in einigen Worten zusammen.

   Bemerkung: Der Bias, der mittlere quadratische Fehler und der Standardfehler des Schätzers T für θ sind auf der Basis von m Schätzungen t_j wie folgt zu schätzen:

$$bias(T,\theta) = \bar{t}_m - \theta, \qquad \bar{t}_m = \frac{1}{m}\sum_{j=1}^{m} t_j$$

$$mse(T) = \frac{1}{m}\sum_{j=1}^{m}(t_j - \theta)^2$$

$$se(T) = \left(\frac{1}{m-1}\sum_{j=1}^{m}(t_j - \bar{t}_m)^2\right)^{1/2},$$

9. Simulierte Konfidenzintervalle für Normalverteilte Zufallsvariablen:

   Führen Sie die folgenden Experimente durch:

   (a) Generieren Sie k Samples der Größe $n = 20$ aus $N(\mu,\sigma)$, $\mu = 2$, $\sigma = 1$. Ermitteln Sie für jedes Sample das entsprechende $(1-\alpha) * 100$-Prozent Konfidenzintervall. Berechnen Sie die relative Überdeckungshäufigkeit

   $$c_k = \text{Anzahl der Überdeckungen von } \mu \text{ durch das Konf.Int.}/k.$$

   Führen Sie die Berechnungen für $k = 20, 50, 100$ und $\alpha = 0.05, 0.01$ durch.

   (b) Variieren Sie nun den Stichprobenumfang n. Generieren Sie jeweils ein Sample der Größe $n = 10, 30, \ldots, 50$ und berechnen Sie die Konfidenzintervalle für μ zum Konfidenzniveau $1 - \alpha$, $\alpha = 0.05$. Geben Sie die errechneten Konfidenzintervalle an.

   (c) Führen Sie 3.1 mit dem Parameter p der Bernoulliverteilung durch. Verwenden Sie für die Konfidenzintervalle die im Vorlesungsskriptum angegebene Approximation. Wählen Sie $p = 0.5$, $\alpha = 0.1$, $k = 1000$ und $n = 50, 100, 300$. Berechnen Sie die relative Überdeckungshäufigkeit c_k. (Vgl. VL-Skriptum Seite 55.) Führen Sie dieses Experiment mehrmals durch und fassen Sie Ihre Resultate zusammen.

10. Mittelwertvergleiche. Prüfen Sie die Hypothesen

$$H_0: \quad \mu_{fvc} = \mu_0 \qquad H_1: \quad \mu_{fvc} \neq \mu_0$$

   für $\mu_0 = 530, 535, \ldots, 570, 575$ zum Signifikanzniveau 0.05%. Welcher Zusammenhang besteht zwischen den Testergebnissen und den von SPSS ausgegebenen Konfidenzintervallen für die Differenz von Mittelwert und Testwert?

11. Mittelwertvergleiche. Ist die Annahme das μ_{fvc} bei den jungen Testpersonen (bis einschließlich 30) größer ist als bei den älteren (ab 31) statistisch zu rechtfertigen. Formulieren Sie dies als Hypothesentest. Dabei soll der Fehler, irrtümlich bei der gegenteiligen Annahme zu bleiben, durch das Signifikanzniveau $\alpha = 0.05\%$ gesteuert werden. Ist der T-Test anwendbar? Gehen Sie in SPSS unter *Analyze–Nonparametric Test–Two Independent Samples.* Wählen Sie den Man-Whitney-U Test aus. (Dieser ist unter allgemeineren Bedingungen als der T-Test einsetzbar.) Versuchen Sie die Ausgabe von SPSS zu interpretieren.

    (Hinweis: Der Man-Whitney-U Test Wurde nicht in der VL besprochen: Die erste Zeile in der von SPSS ausgegebenen Tabelle enthält den Wert der Teststatistik, die vierte den P-Wert.)

12. Seien $X_1, \ldots, X_n \sim N(\mu, \sigma)$, $\sigma = 1$ bekannt. Berechnen Sie die Powerfunktion $\beta(\mu)$ für den folgenden Test:

$$H_0 : \mu \leq 0$$
$$H_1 : \mu > 0,$$

    wobei $H_0 :$ abgelehnt wird, falls $\sqrt{n}\bar{X}_n > z_{1-\alpha}$.

    Dabei bezeichnet $\bar{X}_n$ wie üblich das arithmetische Mittel der $X_1, \ldots, X_n$ und $z_{1-\alpha}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung.

    Simulieren Sie die Funktion $\beta(\mu)$: Generieren Sie für jedes $\mu = -0.2, -0.1, \ldots, 0.8$ $k = 100$ Samples der Größe n aus einer $N(\mu, 1)$-Verteilung. Führen sie für jedes Sample den oben beschriebenen Test durch und schätzen Sie $\beta(\mu)$ durch

$$\hat{\beta}(\mu) = \frac{\sharp\{\text{Tests, die ablehnen}\}}{k}.$$

    Tragen Sie in einem Diagram μ gegen $\beta(\mu)$ und $\hat{\beta}(\mu)$ auf.

    Führen Sie dieses Experiment für $n = 20, 100$ durch und interpretieren Sie Ihr Resultat.

13. Proben von Eisenerz sollen auf ihren Eisengehalt analysiert werden:

62,66	62,87	63,22	63,01	62,10
63,43	63,22	63,57	61,75	63,15
63,08	63,22	63,57	61,75	63,15
61,68	62,45	62,10	62,87	62,87
62,80	62,80	63,01	62,10	63,29
62,80	62,80	63,01	62,10	63,29

    (a) Stellen Sie das Ergebnis graphisch dar. (Histogramm und Boxplot)
    (b) Prüfen Sie, ob eine Normalverteilungsannahme gerechtfertigt ist.

    (*ftp://statistik.tu-graz.ac.at/angstat/data/eisenerz.txt*)

14. Ein Walzwerk liefert Eisenplatten, die von 4 verschiedenen Walzen stammen. Der Verwendungszweck dieser Platten erfordert, daß sie alle die gleiche Dicke besitzen. Zur Untersuchung dieses Merkmals wurden 20 Platten nachgemessen, wobei von jeder der Walzen jeweils 5 dieser Platten stammten. Es ergaben sich die Werte (in mm):

Walzen Nr.	Platten Nr. 1	2	3	4	5
1	9,34	9,38	9,12	9,32	9,28
2	9,67	9,51	9,61	9,52	9,57
3	9,14	9,13	9,06	9,02	9,07
4	9,71	9,75	9,50	9,54	9,55

Es bezeichne x_{ij} die Dicke der j-ten Platte von der i-ten Walze ($1 \leq i \leq 4, 1 \leq j \leq 5$). Unter Annahme, daß die Meßwerte x_{ij} eine Realisierung von unabhängigen, für gleiches i identisch $N(\mu_i, \sigma^2)$-veteilten Zufallsvariablen sind, überprüfe man durch Anwendung eines geeigneten Verfahrens zum Niveau 0,01 die Annahme, daß die mittlere Plattendicke bei allen 4 Walzen gleich ist.
(*ftp://statistik.tu-graz.ac.at/angstat/data/walzen.txt*)

15. Bei einer Studie an Spezialmaschinen in der stahlerzeugenden Industrie wurde untersucht, ob zwischen dem Druck (bar) und der Temperatur (°C) ein linearer Zusammenhang angenommen werden kann. Für diese Untersuchungen wurden folgende Daten ermittelt:

Druck	126	133	134	134	136	137	138	139	141	142	143	147
Temp.	753	809	796	765	812	807	801	795	826	846	817	842

(a) Stellen Sie die Daten in einem Scatterplot dar. Was können Sie erkennen?

(b) Berechnen Sie die Ausgleichsgerade (Regressionsgerade) und geben Sie die Konfidenzintervalle (95%) für die zugehörigen Parameter an. Was können Sie über die Güte der Anpassung sagen?

(c) Überprüfen Sie graphisch, ob alle Vorraussetzungen erfüllt sind.

(*ftp://statistik.tu-graz.ac.at/angstat/data/spezialm.txt*)